Forecasting natural gas consumption using Bagging and modified regularization techniques

Erick Meira*

Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro
Rua Marquês de São Vicente, 225, Ed. Cardeal Leme, 9º andar, Rio de Janeiro 22451-900, BR

Energy, Information Technology and Services Division, Brazilian Agency for Research and Innovation (Finep)
Praia do Flamengo, 200, 9º andar, Rio de Janeiro 22210-030, BR

Fernando Luiz Cyrino Oliveira

Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro
Rua Marquês de São Vicente, 225, Ed. Cardeal Leme, 9º andar, Rio de Janeiro 22451-900, BR

Lilian M. de Menezes

Bayes Business School (formerly Cass), City, University of London
106 Bunhill Row, London EC1Y 8TZ, UK

Abstract

This paper develops a new approach to forecast natural gas consumption via ensembles. It combines Bootstrap Aggregation (Bagging), univariate time series forecasting methods and modified regularization routines. A new variant of Bagging is introduced, which uses Maximum Entropy Bootstrap (MEB) and a modified regularization routine that ensures that the data generating process is kept in the ensemble. Monthly natural gas consumption time series from 18 European countries are considered. A comparative, out-of-sample evaluation is conducted up to 12 steps (a year) ahead, using a comprehensive set of competing forecasting approaches. These range from statistical benchmarks to machine learning methods and state-of-the-art ensembles. Several performance (accuracy) metrics are used, and a sensitivity analysis is undertaken. Overall, the new variant of Bagging is flexible, reliable, and outperforms well-established approaches. Consequently, it is suitable to support decision making in the energy and other sectors.

Keywords: Forecasting, Natural gas demand, Ensembles, Bagging, Regularization

*Corresponding author: erickmeira89@gmail.com; research@erickmeira.com
Email addresses: research@erickmeira.com, cyrino@puc-rio.br, l.demenezes@city.ac.uk

1. Introduction

Natural gas is a strategic energy source in most European countries. Given recent decarbonization initiatives and commitments to phase out electricity generation from coal, natural gas is expected to remain as an important source of energy (IEA, 2021). In absolute terms, the greatest contribution to future global energy demand is expected to come from natural gas, as it is the only fossil fuel whose demand is forecast to globally increase by almost 3 mboe/d from 2019 to 2045 (OPEC, 2020). Although most of this increase is expected to come from developing countries, which have been expanding industrial sectors and electricity consumption, natural gas still plays a major role in Europe. Bastianin et al. (2019) argues, for instance, that the consumption of natural gas in Europe is likely to grow more than that of any other energy source over the period 2015–2050. On the one hand, gross inland consumption of natural gas has considerably increased, following the European Union (EU) Second Strategic Energy Review, which emphasized sustainability, competitiveness and security of supply (CEU, 2008). On the other hand, indigenous natural gas production in Europe is declining because of dwindling reserves (Chen et al., 2019). The International Energy Agency (IEA) estimates that, given the limited reserves of natural gas in the EU and a shift to gas-fired power generation, the EU's dependence on gas imports may reach over 85% by 2030 (IEA, 2020).
As this dependence increases, natural gas markets are becoming increasingly volatile, thus highlighting the risks faced by gas utilities and consumers in Europe. Wood (2016) describes inherent sources of uncertainty, e.g., gas on gas competition (short-term hub prices versus long-term contracts indexed to oil products); third-party access to key infrastructure that increases competition; concerns over carbon pricing and emissions taxes discouraging investment in gas infrastructure; political unrest around the "Southern Corridor" potential pipeline routes, across Turkey and the Balkans; and uncertainties related to technology, decreasing levelized costs of renewables, public pressure and regulation that hinder the future exploitation of gas resources within the EU. In this context, the importance of accurate natural gas demand forecasts for medium term planning can not be overstated. Not only reliable estimates allow businesses to position themselves competitively in markets, but also aid power system operators in balancing electricity supply and demand. Consequently, forecasts of natural consumption are important to maintain gas and electricity reserve margins, and ultimately, to secure energy supply in Europe.

1.1. On forecasting natural gas demand

Several studies have addressed how to forecast natural gas demand. Yet, there is no consensus as to which approaches are more suitable, and little evidence of how methods perform in a wide range of data. Furthermore, the majority of studies are narrow in their geographical coverage.

Sánchez-Úbeda & Berzosa (2007) consider a set of daily industrial end-use natural gas demand series from Spain. Vondráček et al. (2008) use a nonlinear regression model to estimate natural gas consumption of residential and small commercial customers from May 2005 to April 2006 in West Bohemia, Czech Republic. Azadeh et al. (2011) make annual projections of natural gas demand in four Middle Eastern countries from 2008-2015. Taşpınar et al. (2013) consider daily natural gas consumption in the Turkish Sakarya province and test the performance of three methods – Seasonal ARIMA with regressors (SARIMAX), Artificial Neural Networks (ANN) with Multilayer Perceptrons (ANN-MLP) and with a Radial Basis Function layer (ANN-RBF). Potočnik et al. (2014) investigate static and adaptive models when forecasting day-ahead natural gas demand from a local distributor in Croatia. Bai & Li (2016) consider a Support Vector Regression (SVR) approach to forecast daily natural gas consumption in Anqing, China. Panapakidis & Dagoumas (2017) combined Wavelet Transform (WT), Genetic Algorithm (GA), Adaptive Neuro-Fuzzy Inference System (ANFIS) and Feed-Forward Neural Networks (FFNN) to forecast day-ahead natural gas demand in selected Greek distribution points. Özmen et al. (2018), in turn, use Multivariate and Conic Multivariate Adaptive Regression Splines (MARS & CMARS) to forecast day-ahead residential gas consumption in Ankara, Turkey. Beyca et al. (2019) focus on machine learning tools to forecast 12 months-ahead natural gas consumption in the province of Istanbul.

Besides the limited scope in terms of covered regions, a considerable share of previous studies do not provide metrics for out-of-sample forecasting evaluation: most are concerned with scenario projections, rather than providing comparable point forecasts. Error metrics such as the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) are sometimes present, but mainly used for in-sample fit asssessments or model optimization. Out-of-sample evaluation, when present, tend to rely on simple performance metrics in each predicted step, such as the Relative Error (RE). In short, the limited number of competitive benchmarks hinders a clear evaluation of the current state of the literature.

Another point worth noting is that studies that forecast monthly natural gas demand are scarce. Two exceptions are Vondráček et al. (2008) and Beyca et al. (2019), which focus on specific regions: West Bohemia, in the former, and Istanbul, in the latter. However, monthly time spans are of particular relevance, for any strategic decisions in the gas and in the energy sector as a whole, are within this frequency.

Finally, notwithstanding the merits of previous studies, the energy transition towards more flexible and cost effective power systems (Babatunde et al., 2020; Liebensteiner & Wrienz, 2020) and changes in residential consumption patterns (BBC, 2019) calls for adaptive forecasting, i.e., models which

can easily adapt to changes in stylized facts of the time series.

1.2. The growing relevance of ensemble forecasting in the energy sector

Among adaptive forecasting tools, there are hybrid ensemble approaches, which incorporate traditional statistical methods and machine learning techniques, and can deal with different stylized facts and improve forecasting. Zhang et al. (2015) combines Ensemble Empirical Mode Decomposition (EEMD), Least Square Support Vector Machines coupled with Particle Swarm Optimization (LSSVM–PSO) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models to forecast West Texas Intermediate (WTI) crude oil prices. Wang & Wang (2020) put forth a hybrid forecasting model that considers EMD and Gated Recurrent Unit coupled with Stochastic time effective Weights (SW-GRU) to forecast daily futures and spots prices of the WTI and Brent crudes and the Reformulated Blendstock for Oxygenate Blending (RBOB) gasoline. De Oliveira & Cyrino Oliveira (2018) consider a *Bootstrap Aggregation* (*Bagging*) ensemble to forecast monthly electricity consumption in several economies. Agrawal et al. (2019) propose an ensemble of Relevance Vector Machines (RVM) and boosted trees for hour-ahead electricity price forecasting in New England.

Eensembles have also been successful in univariate time series forecasting competitions. For instance, Petropoulos et al. (2018) demonstrate that their Bootstrap Model Combination (BMC) outperforms several benchmarks in the M and M3 Competitions (Makridakis et al., 1982; Makridakis & Hibon, 2000), which respectively comprise 1001 and 3003 time series of different frequencies. Meira et al. (2021b) propose a Pruned BaggedETS algorithm, an ensemble combining *Bagging* with a novel feature selection technique, and obtain promising results on all 98,830 series from the M, M3 and M4 (Makridakis et al., 2019) competitions.

In this context, this paper proposes an approach to ensembles that combines *Bagging* – a class of ensembles approaches –, univariate time series methods and modified regularization to forecast monthly natural gas consumption across 18 European economies. The proposal addresses different stylized facts that may be present in natural gas consumption and related time series, such as nonlinearities, stochastic components (trend, seasonality), heteroscedasticity, and structural breaks. In order to assess forecasting performance, a comparative, out-of-sample analysis is conducted using forecast error metrics and alternative forecasting methods, including traditional benchmarks and state-of-the-art ensembles. Several robustness checks and sensitivity analyses are undertaken. In all, the results are promising.

The next section addresses univariate time series forecasting with ensembles. Section 3 describes the proposed methodology, highlighting its main differences in relation to recently developed ensemble methods for forecasting. Section 4 describes the data and the evaluation setup, and Section 5

summarizes the results and assesses their implications. Finally, Section 6 concludes and suggests directions for future research.

2. A framework for ensemble forecasting methods

Ensembles are supervised learning techniques that follow the concept of Decision Committee Learning (Nock & Gascuel, 1995): committee-members (models) are applied to either a classification or a forecasting task, and their outputs are then combined to create a single forecast. Forecasting with ensembles has 4 main stages: (i) an (optional) data treatment and decomposition, in which the time series is transformed (if necessary) and decomposed into its key components; (ii) the resampling of one or more components multiple times, with subsequent inversion of the initial transformation; (iii) forecasting each series in the resulting ensemble (the original time series and its replicas); and the (iv) combination and pruning stage, in which the forecasts are averaged and some may be removed, when they are unlikely to improve the final result. The flowchart presented in Figure 1 summarizes how these steps are taken for univariate time series forecasting. By considering multiple predictors that are built on replicas of the original data, a random pool (ensemble) of forecasts is formed, and then combined into a single forecast. Hence, the approach allows for the inclusion of different types of uncertainty that may arise when building predictions from data, i.e., data uncertainty, model uncertainty, and parameter uncertainty (Petropoulos et al., 2018). As described in the next subsections, each stage contributes to the performance of an ensemble.

Among ensembles, those which selectively resample from the original data to generate replicas to which a base model is applied have consistently reduced forecasting error. *Bagging* (Breiman, 1996) and *Boosting* (Freund, 1995) algorithms have attracted considerable attention in the forecasting literature. The former generates replicas in parallel, whilst the latter generates them sequentially. Overall, *Bagging* is thought to be more consistent, as it increases the forecasting error less frequently than *Boosting*, whilst the latter may have greater average effect, thus leading to better goodness-of-fit (Webb, 2000). However, much of the gains from *Boosting* seems to be due to overfitting (Quinlan, 1996), which explains its poor forecasting performance on volatile time series. *Bagging* algorithms are particularly useful in forecasting, given their capability to select predictors originated from the forecasting ensemble by means of user-defined techniques, i.e., the practitioner is not restricted to the pre-defined weights of *Boosting*. Hence, *Bagging* is the point of reference to our study.

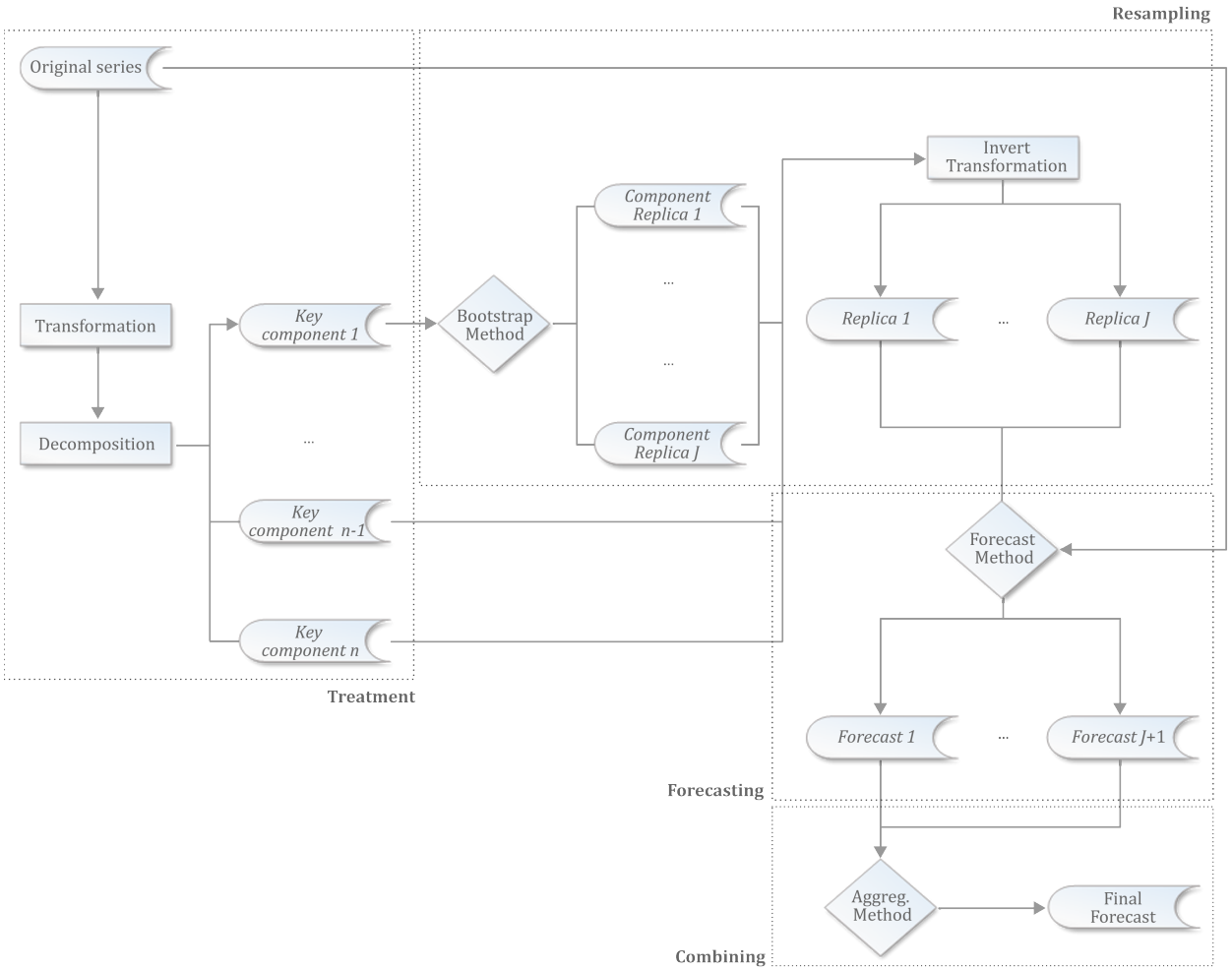

Figure 1: Univariate time series forecasting with ensembles. J stands for the desired number of replicas of the original series – usually 99 in empirical experiments.

2.1. Treatment and decomposition of the time series

In most forecasting ensembles, the time series is initially filtered or smoothed. A common procedure for treating time series data is the Box–Cox (BC) transformation (Box & Cox, 1964), which can simultaneously stabilize the variance, reduce the skewness of the distribution, and ensure that the components of the time series are additive (Petropoulos et al., 2018). Unsurprisingly, several recent studies in forecasting have applied this transformation (Bergmeir et al., 2016; Dantas et al., 2017; De Oliveira & Cyrino Oliveira, 2018; Petropoulos et al., 2018; Dantas & Cyrino Oliveira, 2018; Meira et al., 2021b,a).

After an initial treatment, time series decomposition can be considered, since the estimation error obtained from further aggregating the extrapolated sub-series is expected to be lower than the estimation error for the whole series (Theodosiou, 2011). Two types of decomposition are common in the literature: Seasonal-Trend decomposition using *Loess* (STL) (Cleveland et al., 1990), and Empirical Mode Decomposition (EMD) (Huang et al., 1998). The former consists of six sequential

smoothing operations employing Locally-Weighted Regression (*Loess*) that decompose the series into three additive components: trend, seasonal and remainder (Petropoulos et al., 2020). When compared to other decomposition methods, STL is robust to outliers, can deal with any type of seasonality regardless of the data-frequency, and allows for controlling trend-cycle smoothness (Hyndman & Athanasopoulos, 2021). By contrast, EMD decomposes the time series into a sum of oscillatory Intrinsic Mode Functions (IMFs) that are symmetric with respect to their local zero-mean. The number of extrema and zero-crossings for each IMF are, by definition, equal or allowed to differ at most by one in the whole data. IMFs are more regular and thus easier to forecast.

Most ensembles adopt STL, prior to resampling the time series. STL has also been integrated to hybrid forecasting methods, such as the Bagged.BLD.MBB.ETS by Bergmeir et al. (2016) and the Bootstrap Model Combination of Petropoulos et al. (2018). However, EMD has shown encouraging performance. Recent examples include: the EMD-Holt-Winters *Bagging* approach of Awajan et al. (2018); the Interval Decomposition Ensemble (IDE) of Sun et al. (2018), which combined Bivariate Empirical Mode Decomposition (BEMD), Interval Multilayer Perceptron and Interval Holt's exponential smoothing method (HoltI); and the hybrid EMDHR-SVR-BPNN model of Fan et al. (2020), which integrated empirical mode decomposition, support vector regression and back-propagation neural networks for mid-short-term load forecasting in New South Wales (NSW, Australia).

2.2. Resampling

The rationale behind resampling in ensembles is to generate predictors that share properties of the original data. Resampling can be achieved in many ways (e.g. Monte Carlo simulation, resampling with or without replacement). Two major properties of the time series, however, must be considered while resampling: stationarity and time-dependency. While stationarity may not always be achieved with a single transformation, it is often fulfilled by the remainder from a decomposition method. Concerning the time series structure, variants of the bootstrap algorithm (Efron, 1979) have been used.

In this context, Cordeiro & Neves (2009) proposed a variant of the Sieve Bootstrap (Bühlmann, 1998), which generates replicas of the original series that are independently predicted via exponential smoothing methods, whose forecasts are then combined using the mean or the median. This approach is known as *Boot.EXPOS*, and has outperformed traditional benchmarks from the M3 Forecasting Competition (Makridakis & Hibon, 2000), when forecasting time series with marked seasonal and trend components (mainly quarterly and monthly series, which are common to natural gas and electricity demand data).

Encouraged by the performance of *Boot.EXPOS*, Bergmeir et al. (2016) proposed an alternative in which data are initially treated via a BC transformation, followed by STL decomposition. Replicas for the remainder are then generated via a modified *Moving Blocks Bootstrap* (MBB) algorithm (Künsch, 1989). Once the desired number of replicas is generated, the time series are reconstructed from its structural components and the Box-Cox transformation inverted. Exponential smoothing models are estimated on the original data and each replica, and their point forecasts are then combined using the median. Their approach, which became known as Bagged.BLD.MBB.ETS, outperformed *Boot.EXPOS* and other simple benchmarks, particularly for monthly time series data.

2.3. Forecasting

Following resampling, forecasting models are estimated using the original data and each of its replicas separately. Three families of models are frequently considered at this stage, namely: Neural Networks (NNs), Exponential Smoothing formulations and ARIMA models.

Ensembles of NNs have been used for over 30 years, especially within artificial intelligence, and may include a variety of methods (Adeodato et al., 2011; Barrow & Crone, 2016; Barak & Sadegh, 2016; Khwaja et al., 2017; Szafranek, 2019). They are generally a means to make the most of computing power to address the uncertainty in individual point forecasts. As Rendon-Sanchez & de Menezes (2019) noted, ensembles of NNs have been particularly successful in forecasting short-term electricity demand and were inspirational in the development of combinations of different types of artificial intelligence approaches (Ma et al., 2019) with those obtained from statistical models (Matijaš et al., 2013).

Exponential smoothing, in turn, are methods that attribute exponentially decreasing weights for past data, i.e., recent observations are given greater weight in forecasting than older ones. Although the basic specifications date from the seminal works of Holt (1957, reprinted 2004) and Winters (1960), exponential smoothing methods remain widely applied, mainly due to their simplicity and transparency, but also due to their adaptability to changes in the time series (Goodwin, 2010). In addition, exponential smoothing has a theoretical foundation in state space modelling, which allows for straightforward implementations in statistical packages (Hyndman et al., 2002, 2008; Hyndman & Athanasopoulos, 2021). Exponential smoothing models defined within this state-space framework are often referred to as ETS, an acronym to 'ExponenTial Smoothing' or 'Error, Trend and Seasonal', the components of the time series which vary across formulations, i.e., additive, additive damped, multiplicative, or multiplicative damped.

ARIMA (Autoregressive, Integrated, Moving Average) formulations, in turn, stems from the Box-Jenkins (1970) family of models. They are similar to exponential smoothing, as they can model

trends and seasonal patterns and be automated, but are based on autocorrelation and partial autocorrelation functions of the time series rather than a structural view of the time series (level, trend and seasonality).

2.4. Combining

The final stage in ensemble consists of combining (aggregating) forecasts. Two combinations are predominant: the mean (or equal weights combination) – the simplest, yet, a robust approach (Stock & Watson, 2004) – and the median, which may dilute the effects of occasional poor forecasts. While Cordeiro & Neves (2009) in their ensemble, *Boot.EXPOS*, use both, Bergmeir et al. (2016) adopt the median in their Bagged.BLD.MBB.ETS. By contrast, Petropoulos et al. (2018) propose a more sophisticated combination: Bootstrap Model Combination (BMC). As in Bagged.BLD.MBB.ETS, replicas are originated by resampling the remainder from a STL decomposition and are predicted using exponential smoothing specifications. However, replicas in BMC drive the selection of the best-fit model, since forecasts are combined using weights reflecting the frequency that the selected model specifications were identified as best-fit on the pool of replicas. Considering all series from two forecast competitions, M (Makridakis et al., 1982) and M3 (Makridakis & Hibon, 2000), the BMC outperformed the Bagged.BLD.MBB.ETS.

Dantas & Cyrino Oliveira (2018) combine *Bagging* with clustering. Their Bagged.Cluster.ETS aims at reducing covariance in the ensemble. Partitioning Around the Medoids (PAM) (Kaufman & Rousseeuw, 1987) is used to identify clusters of similar forecasts, and then forecasts from each cluster are selected in order to create a smaller subset of forecasts with reduced error-variance to be combined using the median. This method was evaluated on series from the M3 and CIF 2016 competitions, and outperformed several benchmarks, including other *Bagging* approaches.

2.5. Limitations in ensembles for time series forecasting

Table 1 summarizes the main features of established *Bagging* algorithms for forecasting, highlighting their strategies in each step of algorithm and indicating their main limitations. Overall, criticisms of *Bagging* ensembles concern the resampling and combining stages. Resampling has been mostly conducted via the modified Moving Blocks Bootstrap (MBB) algorithm proposed by Bergmeir et al. (2016) – see, for instance, Dantas et al. (2017); De Oliveira & Cyrino Oliveira (2018); Petropoulos et al. (2018); Dantas & Cyrino Oliveira (2018) and Meira et al. (2021b). However, MBB is very sensitive to the choice of the block size, for which there is no consensus on what would be optimal. In addition, MBB, like most bootstrapping approaches, repeats some original values and does not use many others, and thus values in the neighborhood of observed points in the time series may not

be included in a replica. It also restricts values to lie in the closed interval $[min(x_t), max(x_t)]$, where x_t is the original series. Hence, alternatives to the MBB have been proposed. As previously outlined, Cordeiro & Neves (2009) consider a variant of the sieve bootstrap of Bühlmann (1998) where Autoregressive (AR) models are applied to the residuals of an exponential smoothing formulation on the original series. The new residuals, after AR fitting, are then considered for resampling with replacement. Although the authors ensure that characteristics of the original time series are kept, the quality of the resampling is depends on the quality of the selected exponential smoothing model. Petropoulos et al. (2018) consider the Circular Blocks Bootstrap (CBB) and the Linear Process Bootstrap (LPB). The former has the same restrictions as the MBB, although it is theoretically superior since the time series are 'wrapped' in a circle before resampling takes place, ensuring that the first and last observations are not subsampled. LPB involves estimating the autocovariance matrix of the original series and pre-whitening the noise with the estimated matrix; after which, it generates replicas from the pre-whitened noise and postcolors the noise with the autocovariance matrix. The findings of Petropoulos et al. (2018) suggest similar results from the three resampling algorithms. De Oliveira & Cyrino Oliveira (2018) proposed a resampling scheme, named after Remainder Sieve Bootstrap (RSB), which fits Autoregressive, Moving Average (ARMA) models first and resamples from the residuals. Just as the LPB, RSB allows for greater coverage in replicas. Nonetheless, some theoretical limitations remain, such as not satisfying the ergodic theorem (see the next section for further details).

Concerning limitations in the combining/aggregation stage, most *Bagging* approaches consider the mean or the median for forecast aggregation – see, for instance, Cordeiro & Neves (2009), Bergmeir et al. (2016) and De Oliveira & Cyrino Oliveira (2018). When more sophisticated approaches are adopted, only feature selection is achieved. Bagged.Cluster.ETS (Dantas & Cyrino Oliveira, 2018), for instance, considers PAM clustering to drive feature selection of the forecasts, but the aggregation of the remaining forecasts in the ensemble remains via the median. The same holds for the Pruned BaggedETS approaches of Meira et al. (2021b) and Meira et al. (2021a): pruning can effectively conduct feature selection via outlier detection in the prediction intervals of the forecasts; but these methods still rely on traditional averaging. The BMC of Petropoulos et al. (2018) can imply variable weights, but it is restricted to exponential smoothing and may generate considerably large forecasts when applied to series with structural breaks or outliers, as depicted in Meira et al. (2021b).

Method(s) & Source	Treatment & Decomposition	Resampling algorithm(s)	Forecasting method(s)	Combining strategies	Limitations
Boot.EXPOS (Cordeiro & Neves, 2009)	Four selected exponential smoothing models[1] plus AR model fit to residuals	Sieve Bootstrap on the residuals	Four selected exponential smoothing models[1]	Mean, Median	Resampling dependent on the selected exponential smoothing model; Limited selection of methods for forecasting; No feature selection nor combination
Bagged.BLD.-MBB.ETS (Bergmeir et al., 2016)	BC transformation & STL Decomposition	MBB	ETS	Median	MBB sensitive to block size; Replicas must lie in the interval $[min(x_t), max(x_t)]$; No feature selection nor combination
RSB & MBB BaggedETS & BaggedARIMA (De Oliveira & Cyrino Oliveira, 2018)	BC transformation & STL Decomposition	MBB, RSB	ETS, ARIMA	Mean, Median	No feature selection nor combination
Bootstrap Model Combination (BMC) (Petropoulos et al., 2018)	BC transformation & STL Decomposition	CBB, LPB, MBB	ETS, ARIMA	Weighted average, weights reflect the frequency of model forms during estimation	May generate large forecasts when applied to series with notable structural breaks or outliers
Bagged.-Cluster.ETS (Dantas & Cyrino Oliveira, 2018)	BC transformation & STL Decomposition	MBB	ETS	Partitioning Around Medoids (PAM) for feature selection, followed by median aggregation	MBB-related (block size & closed interval); Only feature selection is achieved (no variable weighting); Computing intensive
Pruned BaggedETS & Pruned Bagged-TreatedETS (Meira et al., 2021b; 2021a)	BC transformation & STL Decomposition	MBB (both), CBB and LPB in Meira et al. (2021a)	ETS (both), TreatedETS in Meira et al. (2021b)	Pruning for feature selection, followed by median aggregation	Only feature selection is achieved (no variable weighting);

Table 1: Main features and limitations of established *Bagging* ensembles for time series forecasting. Notes: [1]The four methods considered by Cordeiro & Neves (2009) are: Single exponential smoothing, Holt's linear trend, Additive Holt–Winters and Multiplicative Holt–Winters.

3. Proposed ensemble

Our proposal covers the same core ideas of *Bagging* for forecasting, but significantly differs from previous proposals in two ways. First, replicas are generated via the Maximum Entropy Bootstrap (MEB), using a seven-step algorithm that satisfies the ergodic theorem by ensuring that the grand mean of all ensembles is close to the original sample mean (Vinod, 2004). Secondly, modified regularization (M-Ridge and M-LASSO) is used to aggregate (combine) the forecasts. The modified regularization procedures differ from traditional Ridge and LASSO routines because they ensure that the data generating process of the forecasts in the ensemble is kept. In the next subsections, each stage is described.

3.1. Data treatment and resampling

The first part of our approach is akin to established *Bagging* forecasting approaches, since it involves generating replicas for the remainder component of an STL decomposition applied to a Box–Cox (BC) transformed time series (Bergmeir et al., 2016; Dantas et al., 2017; De Oliveira & Cyrino Oliveira, 2018; Petropoulos et al., 2018; Dantas & Cyrino Oliveira, 2018; Meira et al., 2021b,a). However, instead of using the MBB algorithm to replicate the remainder, a Maximum Entropy Bootstrap (MEB) routine is adopted, so that ensembles are created from a density distribution that satisfies the maximum entropy principle (Vinod & López-de-Lacalle, 2009). To the best of our knowledge, this method has not been adopted within *Bagging*.

Maximum Entropy Bootstrap (MEB) was devised by Vinod (2004) as a resampling procedure for non-stationary time series or when stationarity is difficult to ascertain. Replicas of a time series are generated according to an algorithm designed to ensure that the grand mean of all ensembles is close to the original sample mean. That is, for a time series x_t of size T, the following steps are performed:

1. Sort the data in increasing order to create order statistics $x_{(t)}$ and store the ordering index vector;
2. Compute the intermediate points from the order statistics: $z_{(t)} = \left[x_{(t)} - x_{(t-1)}\right]/2$, $t = 2, 3, \cdots, T-1$;
3. Calculate the trimmed mean (m_{trm}) of the deviations $x_{(t)} - x_{(t-1)}$ among all consecutive observations. In addition, compute the lower and upper limits of the density distribution function, $z_0 = x_{(1)} - m_{trm}$ and $z_T = x_{(T)} + m_{trm}$, respectively;
4. Construct the maximum entropy density function with the z values as limiting points. The density is built by joining uniform distribution intervals of equal probability. The uniform

densities are also designed to satisfy the mean-preserving constraint. To that end, the interval means for the uniform density, m_t, must satisfy the following relations:

$$\begin{aligned} m_1 &= 0.75x_{(1)} + 0.25x_{(2)} \\ m_k &= 0.25x(k-1) + 0.50x_{(k)} + 0.25x_{(k+1)}, k = 2, \cdots, T-1 \\ m_T &= 0.25x_{(T-1)} + 0.75x_{(T)} \end{aligned} \tag{1}$$

5. Inverse transform sampling: generate T random numbers from the $[0, 1]$ uniform interval, compute sample quantiles of the ME density at those points and sort in ascending order;
6. Reorder the the sorted sample quantiles by using the ordering index of step 1. This recovers the time dependence relationships of the originally observed data;
7. Repeat 1–6 until the desired number of replications (J) is achieved.

MEB is attractive because it retains the basic shape and time-dependence structure of the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) of the original time series in its replicas, without resorting to shape-destroying transformations such as detrending or differencing to achieve stationarity (Vinod & López-de-Lacalle, 2009). Besides avoiding transformations to achieve stationarity, according to Vinod (2006), MEB procedure avoids other limitations of standard bootstrapping, which are:

(i) Replicas obtained from shuffling with replacement repeat some original values while excludes many others. They never admit nearby data values in a resample. *A priori*, there is no reason to believe that values near the observed x_t are impossible;

(ii) Replicas must lie in the closed interval $[min(x_t), max(x_t)]$. Since the observed range is random, we cannot rule out somewhat smaller or larger x_t.[1];

(iii) Traditional bootstrap involve shuffling x_t in a way that serial correlation can be lost. Hence, it is impossible to generate a large number of sensibly distinct replicas in a traditional bootstrap.

In addition, MEB is of straightforward implementation and is available in different statistical packages[2]. Hence, the procedure has been effectively applied while investigating associations between energy consumption and economic health in Turkey (Yalta, 2011) and as an auxiliary technique when estimating air temperature quantiles in Central Europe (Barbosa et al., 2011).

[1] Note that the third step of the MEB algorithm implies a less restrictive/wider range $[z_0, z_T]$

[2] In R, MEB can be implemented using the `meboot()` function of the *meboot* package (Vinod & López-de-Lacalle, 2009). Following previous studies (Vinod, 2004, 2006), we set the trimming proportion to 10% by adding trim = 0.10.

Notwithstanding these observations, we are unaware of previous applications of MEB in time series forecasting.

3.2. Forecast generation

After obtaining the desired number of replicas, a forecasting model is estimated for the original data and each of its replicas separately. The models are then used to generate forecasts for the desired forecast horizon. Hence, an ensemble of forecasts is formed, with the number of forecasts equal to the number of replicas generated in the previous step plus one, since one forecasting model is also selected for the original time series. Two widely used families of univariate forecasting models, ETS and SARIMA, are considered.

ETS stands for a finite set of state space based exponential smoothing models, which can be obtained through variations in the combination of the components of a time series. The possible combinations for the trend and seasonal components are depicted in Table 2. In addition, since the error term can be either additive or multiplicative, a total of 30 different ETS models can be achieved (Hyndman et al., 2002). The ETS algorithm fits all variants to the time series. The input is a vector formed by the original data values in a time series format. The output is a model (together with the optimal parameters) consisting of three terms: error, trend, and seasonality. Depending on the formulation, the number of smoothing hyperparameters may include one or more constants (e.g. α , β, γ and ϕ). Each model consists of a measurement equation that describes the observed data, and some state equations that describe how the unobserved components or states (level, trend, seasonal) change over time.

Components	Seasonal		
Trend	None (N)	Additive (A)	Multiplicative (M)
None (N)	N, N	N, A	N, M
Additive (A)	A, N	A, A	A, M
Additive Damped (A_d)	A_d, N	A_d, A	A_d, M
Multiplicative (M)	M, N	M, A	M, M
Multiplicative Damped (M_d)	M_d, N	M_d, A	M_d, M

Table 2: Possible combinations of seasonal and trend components under the ETS state space framework.

After fitting all the formulations to the time series, model selection is performed by choosing the ETS combination that leads to the lowest value of the Akaike Information Criterion with corrections (AICc) (Sugiura, 1978), as commonly adopted in the literature. Finally, the model is used to generate the forecasts for the forecast horizon.

Concerning the practical implementation of the ETS, the optimal model (and its respective hyper-parameters) is identified for each time series using the `ets()` function from the *forecast* package (Hyndman et al., 2021). Forecasts for each selected model are then computed for the desired forecast horizon using the `forecast()` function, in the *forecast* package.

SARIMA models (Box & Jenkins, 1970) are an alternative and complementary approach to exponential smoothing methods. While the latter are based on a structural view of the data (the level, trend, and seasonal components of a time series), SARIMA models focus on serial correlations. In practice, SARIMA tends to outperform exponential smoothing methods for longer, more stable data (De Oliveira & Cyrino Oliveira, 2018). SARIMA models are denoted by $SARIMA(p,d,q) \times (P,D,Q)_S$ and can be written as follows:

$$\nabla_S^D \nabla^d \phi(B)\, \Phi\left(B^S\right) y_t = c + \theta(B)\, \Theta\left(B^S\right) \varepsilon_t \tag{2}$$

where:

- ε_t is a white noise process with mean zero and variance σ^2;
- B is the backward shift operator (eg. $By_t = y_{t-1}$);
- c is a drift parameter. If $c \neq 0$, there is an implied polynomial of order d in the forecast function;
- p, d, and q are non-negative integers respectively referring to the order of the autoregressive model, the degree of differencing, and the order of the moving-average model;
- S refers to the number of periods in each season;
- the uppercase P, D and Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model;
- $\phi(B)$ and $\Phi\left(B^S\right)$ are the non-seasonal and seasonal autoregressive polynomials;
- $\theta(B)$ and $\Theta\left(B^S\right)$ are the non-seasonal and seasonal moving-average polynomials;
- ∇^d and ∇_S^D are the non-seasonal and seasonal differencing operators, respectively.

The SARIMA model selection and forecasting routine employed in the manuscript follows a variation of the Hyndman-Khandakar algorithm (Hyndman et al., 2021). As in ETS, the approach first selects the best fit SARIMA model and then uses this model to generate the forecasts for the desired forecast horizon (number of steps ahead). Model selection uses unit root tests to infer the order of integration of the time series and Maximum Likelihood Estimation (MLE) coupled with the minimization of the AICc (Sugiura, 1978) for lag selection on the stationary part of the model (the

autoregressive-moving average process). The stepwise search process for the non-seasonal ARIMA is detailed below:

Algorithm 1 Hyndman-Khandakar algorithm for ARIMA model selection

Step 1 **Differencing and stationarity (I(d) order selection)**
- Start with $d = 0$ (original time series)
- Test for stationarity using unit root tests (KPSS as default)
- For KPSS, if $p < 0.05$, differencing is required ($d = 1$). Else, $d = 0$
  - If $d = 1$, test once again for stationarity using KPSS
    - If $p < 0.05$, differencing is once again required ($d = 2$)
    - Else $d = 1$

Step 2 **AR(p) and MA(q) lag order selection**
- The values of p and q are chosen by minimising the AICc
- Rather than considering every possible combination of p and q,
- the algorithm uses a stepwise search to traverse the model space

Step 2a **Fit four initial models**
- ARIMA$(0, d, 0)$
- ARIMA$(2, d, 2)$
- ARIMA$(1, d, 0)$
- ARIMA$(0, d, 1)$
- A drift constant (c) is included unless $d = 2$
  - If $d \leq 1$, an additional model is also fitted: ARIMA$(0, d, 0)$ without drift

Step 2b **The model with the lowest AICc in 2a is set as the 'current model'**

Step 2c **Variations of the current model are considered**
- Vary p and q from the current model by ± 1

Step 2d **Include/exclude c from the current model**
- The best model considered so far becomes the new 'current model'

Repeat Step 2c until no lower AICc can be found

Notes: KPSS stands for the Kwiatkowski et al. (1992) unit root test.

The modelling procedure (lag/order selection) for Seasonal ARIMA (SARIMA) models is similar to the described above, except that seasonal AR and MA terms also need to be selected.
In practice, the best SARIMA model for each series is obtained using the `auto.arima()` function from the *forecast* package in R. After identifying the best-fit SARIMA model, forecasts are computed for a desired number of steps-ahead using the `forecast()` function.

3.3. Combining forecasts via traditional and modified regularization

In contrast to most Bagging approaches, rather than taking the mean or median of forecasts, regularization routines assign weights for each forecast in the ensemble via multiple regression. The aim is to significantly reduce the variance of the final forecasting error, though at the cost of introducing some bias. This is an approach which can improve the predictive performance of the

model when: (i) there are many predictors; and/or (ii) the predictors are highly correlated with each other. Hence, regularization should improve ensembles.

Regularization can be viewed as an infinite set of techniques, for which two extreme cases are frequently used in multiple regressions: Ridge Regression (Hoerl & Kennard, 1970) and Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). In both cases, the traditional ordinary least squares loss function is augmented, so that the sum of squared residuals is minimized and large parameter estimates are penalized[3]. Let n be the number of observations of the response variable, Y, represented by a linear combination of m predictor variables, X; and a normally distributed error with σ^2 variance. Under Ridge, the loss function is:

$$L_{Ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^{'}\ \hat{\beta})^2 + \lambda \sum_{j=1}^{m} \hat{\beta}_j^2 \tag{3}$$

where λ is the regularization penalty parameter. Minimizing the above formula gives the Ridge regression estimates $\hat{\beta}_{Ridge} = (X^{'}X + \lambda I)^{-1}(X'Y)$, where I stands for the identity matrix. By incorporating the regularization coefficient in the formulas for bias and variance, we obtain:

$$\begin{aligned} Bias(\hat{\beta}_{Ridge}) &= [(X'X + \lambda I)^{-1} - (X'X)^{-1}]X'X\beta \\ Var(\hat{\beta}_{Ridge}) &= \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1} \end{aligned} \tag{4}$$

From eq. 4, we observe that as λ becomes larger, the variance decreases, and the bias increases. Hence, there is a trade-off to be considered, for which there are basically two strategies. A traditional approach would be to choose the λ that minimizes an information criterion. An alternative is to perform cross-validation and select the value of λ that minimizes the cross-validated sum of squared residuals (or some other measure). The former emphasizes goodness-of-fit and the relative impact of exogenous inputs in the variable of interest, while the latter is focused on predictive performance. Here, this second strategy is adopted, by choosing a set of P values of λ to test, splitting the dataset into K folds, and selecting the optimal λ according to Algorithm 2.

Our implementation uses the `cv.glmnet()` function from the *glmnet* package in R (Friedman et al., 2010) and considers $K = 10$ cross-validation folds and $P = 1000$ possible lambda values, whose sequence is defined by the own function. The value λ_{opt}, which minimizes the average sum of squared residuals, is obtained using a validation set of the same size of the test set.

In the case of LASSO regularization, the loss function is:

[3]There are also the elastic-net models, which are half-way house between the Ridge and the LASSO formulations, obtained by varying the α, the elastic-net penalty parameter over the range of 0 (Ridge) – 1 (LASSO) – see Friedman et al. (2010) for further details. We have considered several versions of these models, but they did not offer improvements over Ridge or LASSO.

$$L_{LASSO}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^{'}\ \hat{\beta})^2 + \lambda \sum_{j=1}^{m} \left|\hat{\beta}_j\right| \tag{5}$$

LASSO adds a penalty for non-zero coefficients but, unlike Ridge, which penalizes the sum of squared coefficients (L2 penalty), LASSO penalizes the sum of their absolute values (L1 penalty). Consequently, for high values of λ, many coefficients become zero under LASSO, which is never the case when using Ridge.

Algorithm 2 Choice of lambda

procedure cross-validation(P = nlambda, K = nfolds)
  for p in 1 to P **do**
    for k in 1 to K **do**
      keep fold k as hold-out data
      use the remaining folds and λ=λ_p to estimate $\hat{\beta}_{Ridge}$
      predict hold-out data: $y_{test,k} = X_{test,k}\ \hat{\beta}_{Ridge}$
      compute the sum of squared residuals: $SSR_k = \|y - y_{test,k}\|^2$
    end for k
    average SSR over the folds: $SSR_p = 1/k \sum_{k=1}^{K} SSR_k$
  end for p
  choose optimal λ value: $\lambda_{opt} = \underset{p}{\operatorname{argmin}}\ SSR_p$
end procedure

where $\|\quad\|^2$ is the quadratic norm.

3.3.1. Modified regularization

Besides traditional Ridge and LASSO regressions, we consider a Modified Regularization, that generates forecasts once, for the period comprising both the validation and combination (test) phases. The rationale behind is that, by conducting validation and combination in the same set of forecasts, the data generating process of the forecasts is kept. This is a subtle difference that can significantly improve accuracy, as can be seen in Section 5.

Figure 2 illustrates the differences between the traditional regularization and our approach. In the former, a training set is used to generate a set of K steps ahead forecasts for the period comprising the validation set, which is first used to compute the optimal regularization parameters for the set of the forecasts and then added to the train set before estimating the final models. This implies that the models that are used to generate the set of forecasts for the validation set are not the same models that are used to generate the final set of forecasts for the test set period, which are combined via regularization to produce the final forecast. In our Modified regularization, forecasts for each replica are computed up to $2K$ steps-ahead: the first K steps relate to the validation set, and are

used to optimize the weights of the regularized model; the last K half and the optimal weights obtained in the validation set (first half) are then used for combining forecasts.

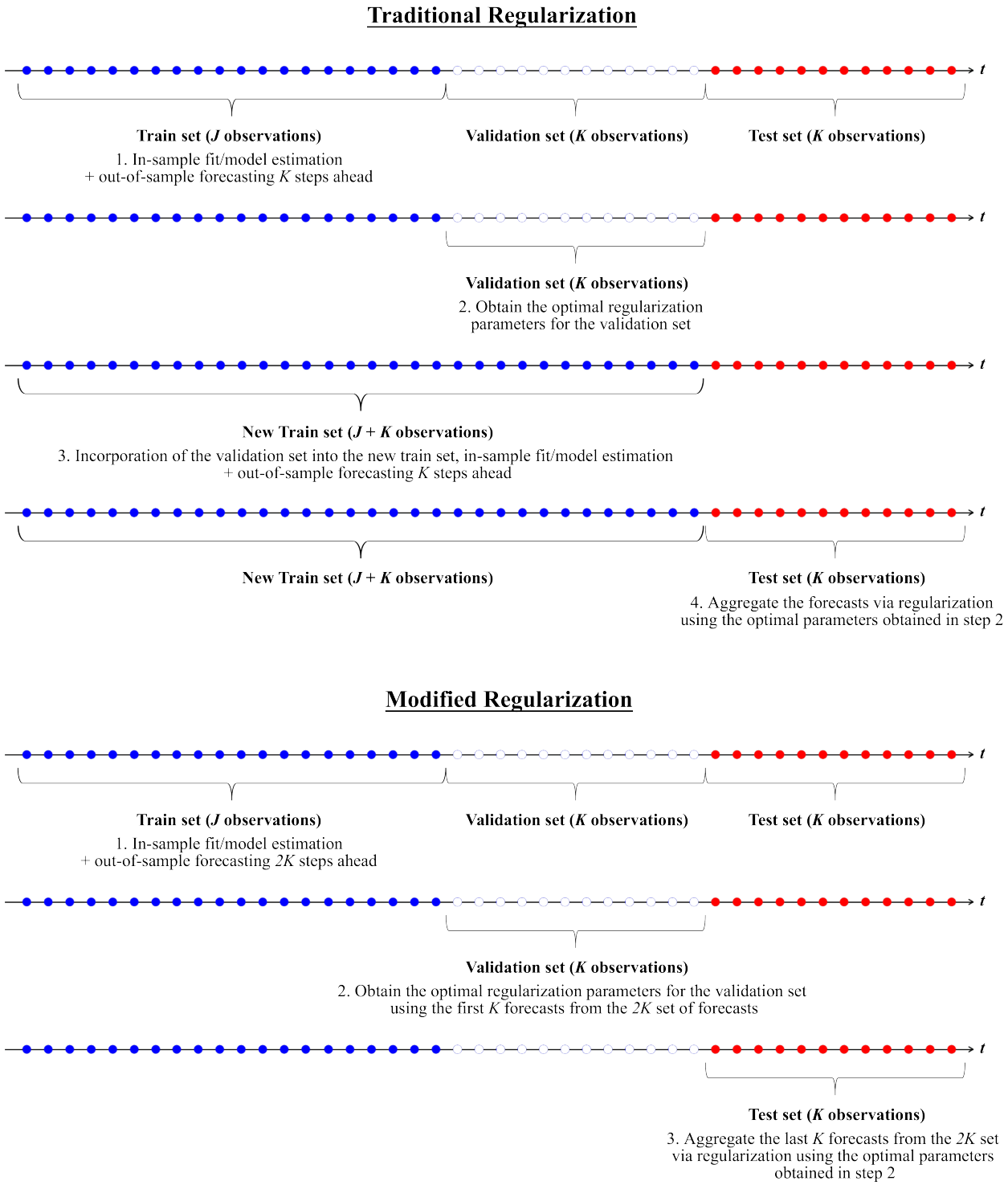

Figure 2: Comparing Traditional and Modified Regularization.

4. Data, comparators and evaluation setup

Monthly data of Gross Inland Natural Gas Consumption (in terajoules, TJ) across several European economies were collected from the Statistical Office of the European Union database (EUROSTAT, 2020). The data span 18 countries[4] from January 2008 to December 2019 (the last official date available for all involved European countries). Observations from January 2008 to December 2018 comprise the training set for benchmark forecasting methods and ensemble approaches using the median for aggregation. When employing regularization, a validation set is included between January 2018 and December 2018, in which weights are assigned to each forecast in the selected ensemble. The test set comprises the last 12 observations: January 2019 – December 2019. Figure 3 depicts the training set of the original series, and descriptive statistics are provided in Appendix A. As can be noted, the selected time series differ considerably in their time-series behavior and shapes, thus highlighting the challenge faced by forecasters in the energy sector.

Forecasts of our proposed ensembles are compared with those from several univariate forecasting methods, which are summarized in Table 3. The first set of comparators are traditional, statistical time series forecasting methods, i.e.:

- The auto, state space exponential smoothing approach (Hyndman et al., 2002), as outlined in Section 3.2, applied to the original time series. More specifically, the `ets()` function from the *forecast* package in R is automatically used for model selection and the `forecast()` function is used to generate the point forecasts for the desired forecast lead time;
- The ARIMA formulation selection algorithm (Hyndman & Khandakar, 2008) and the `forecast()` function are adopted as described in Section 3.2;
- A three parameter Additive Holt-Winters model (Holt, 1957, reprinted 2004; Winters, 1960), applied to the original series via the `hw()` function from the *forecast* package in R;
- A three parameter Multiplicative Holt-Winters model. This model uses the same function of its additive version (`hw()` function), with the seasonal argument set to ''multiplicative";
- The Theta method (Assimakopoulos & Nikolopoulos, 2000), which is akin to a simple exponential smoothing with drift, but with a particular restriction for this last component. This method is known for its predictive performance on monthly series and microeconomic data (Makridakis & Hibon, 2000).

[4]The following countries are included: Austria (AT); Belgium (BE); Czech Republic (CZ); Denmark (DK); Finland (FI); France (FR); Germany (DE); Ireland (IE); Italy (IT); Latvia (LV); Luxembourg (LU); Netherlands (NL); Poland (PL); Portugal (PT); Slovakia (SK); Slovenia (SI); Spain (ES); and United Kingdom (UK).

Figure 3: Gross inland natural gas consumption in terajoules (TJ). Train set observations. Source: EUROSTAT (2020).

Method	Implementation / Source	Short description
Univariate time series forecasting benchmarks		
ETS	R *forecast* package `ets()` function	Automatic Error, Trend and Seasonality specification
ARIMA	R *forecast* package `auto.arima()` function	Automatically-selected (S)ARIMA model
Additive HW	R *forecast* package `hw()` function[1]	Three parameter Additive Holt-Winters method
Multiplicative HW	R *forecast* package `hw()` function[2]	Three parameter Multiplicative Holt-Winters method
Theta	R *forecast* package `theta()` function	Simple exponential smoothing with drift, with a particular drift restriction
Machine learning methods		
ANN	R *forecast* package `nnetar()` function	Single hidden layer, feed-forward neural network with lagged inputs of the time series
BC-ANN	R *forecast* package `BoxCox()` & `nnetar()` functions	The above ANN model with prior Box-Cox transformation
SVR	R *FSelector* package `cfs()` & R *e1071* package `svm()` functions	Support Vector Regression with the set of lagged values selected using the CFS algorithm
RSSA	R *Rssa* package `ssa()` & `rforecast()` functions	Recurrent Singular Spectrum Analysis
VSSA	R *Rssa* package `ssa()` & `vforecast()` functions	Vector Singular Spectrum Analysis
Alternative Bagging approaches		
Bagged.BLD.MBB.ETS (BaggedETS)	Bergmeir et al. (2016)	Median aggregation of ETS forecasts built on bootstraps of the original series.[3]
Bagged.Cluster.ETS	Dantas & Cyrino Oliveira (2018)	Median aggregation of specific BaggedETS forecasts, selected via a Partitioning Around Medoids (PAM) algorithm.
BMC	Petropoulos et al. (2018)	Bootstrap model combination of specific BaggedETS forecasts[3].

Table 3: Comparators. Notes: `ets()` and `auto.arima()` are used for model selection. The `forecast()` function must be used on the output to generate the forecasts. [1]Set seasonal argument to "additive"; [2]Set seasonal argument to "multiplicative". [3]See Section 2.4 for details.

The second set of competing approaches concerns univariate machine learning methods:

- A single hidden layer, feed-forward Artificial Neural Networks (ANN) model (Rumelhart et al., 1985). This model uses as inputs lagged values of the original time series and is capable of addressing complex nonlinear behavior;
- A feed-forward Artificial Neural Network model with prior Box-Cox (Box & Cox, 1964) transformation (BC-ANN);
- A univariate Support Vector Regression (Vapnik, 1995) backed by the Correlation-based feature selection (CFS) algorithm (Hall, 1999) to select the best subset of lags for prediction. SVR learns from the training data and forms complex non-linear decision boundaries. CFS ranks identified attributes according to a heuristic evaluation function, and assumes that irrelevant features should be ignored;
- The Recurrent variation of the univariate Singular Spectrum Analysis (SSA) for forecasting (RSSA). In brief, SSA is a decomposition-reconstruction method that filters the noise and forecast the signal of an underlying time series according to multiple steps (Embedding, Singular Value Decomposition, Grouping and Diagonal Averaging) (Golyandina et al., 2001);
- The Vector variation of the univariate SSA for forecasting (VSSA).

The third set of comparators are *Bagging* algorithms that demonstrated promising results when forecasting monthly time series from international forecasting competitions (Makridakis & Hibon, 2000; Makridakis et al., 2019). Particularly, we consider: the Bagged.BLD.MBB.ETS by Bergmeir et al. (2016) (henceforth referred to as 'BaggedETS', for simplicity); the Bagged.Cluster.ETS method of Dantas & Cyrino Oliveira (2018); and the Bootstrap Model Combination (BMC) of Petropoulos et al. (2018). For details on these methods, see Section 2.4.

Implementation is conducted using the R programming language (R Core Team, 2021) and related packages. We used R version 4.0.2 (2020-06-22) and forecast package version 8.12 for ETS and ARIMA modelling. MEB resampling and traditional regularization were conducted using packages meboot (1.4-8) and glmnet (4.0-2), respectively. Furthermore, a parallel implementation is adopted, using the following packages: *doSNOW* (1.0.18), *foreach* (1.5.0) and *snow* (0.4–3). 99 replicas are generated for each ensemble. To facilitate eplication of our results, all resampling procedures use the same random seed, set to 123 using the `set.seed()` function in R. Block size for MBB resampling in these cases comprised 24 observations, following the same guidelines as established *Bagging* approaches. Pretreatment for all ensemble methods involved using BC transformation and STL decomposition prior to resampling.

To gauge the overall accuracy of the forecasts, the results are summarized according to the mean across all time series of each metric in Table 4. Forecasting performance is also assessed by considering the distribution (boxplots) of each metric obtained when methods were individually applied to each time series. These plots enable an assessment of which time series are more difficult to forecast, as well as which methods vary their performance considerably across time series.

Metric	Formula	Unit of measurement
Root Mean Squared Error (RMSE)	$\sqrt{\frac{\sum_{t=1}^{h}(Y_t-\hat{Y}_t)^2}{h}}$	Same as the original series
Mean Absolute Percentage Error (MAPE)	$\frac{100}{h}\sum_{t=1}^{h}\frac{\left\|Y_t-\hat{Y}_t\right\|}{\left\|\hat{Y}_t\right\|}$	Percentage points (%)
Symmetric Mean Absolute Percentage Error (sMAPE)	$\frac{200}{h}\sum_{t=1}^{h}\frac{\left\|Y_t-\hat{Y}_t\right\|}{\|Y_t\|+\left\|\hat{Y}_t\right\|}$	Percentage points (%)
Mean Absolute Scaled Error (MASE)	$\frac{1}{h}\frac{\sum_{t=1}^{h}\left\|Y_t-\hat{Y}_t\right\|}{\frac{1}{n-m}\sum_{t=m+1}^{n}\|Y_t-Y_{t-m}\|}$	Dimensionless

Table 4: Evaluation metrics. Notes: Y_t e $\hat{Y}_t$ are the real (actual) and forecasted values of the time series, respectively; h is the forecasting horizon (number of forecasting steps ahead); m is the seasonal period.

The choice of metrics (specially MAPE and sMAPE) was mainly to allow comparability with published results. In addition, sMAPE and MASE are the official evaluation metrics for point forecasts in the M4 Competition (Makridakis et al., 2018). MASE is a scale-free metric devised by Hyndman & Koehler (2006). As for RMSE, although averaging its values across multiple series is unusual, it provides an estimate of how much energy (in TJ) might have been "saved" by opting for a more accurate forecasting approach, and is therefore relevant in the context of this study.

Multiple Comparisons with the Best (MCB) approach is also adopted. Essentially, MCB tests whether the average (across time series) rank of each method is significantly different, from the statistical viewpoint, from those of other methods. If the intervals of two methods do not overlap, performances are judged to be statistically different. MCB has been used extensively in the forecasting literature (e.g. Koning et al. (2005); Petropoulos et al. (2019); Spiliotis et al. (2019); Meira et al. (2021b)).

5. Results and Discussion

5.1. Aggregate results and distribution of evaluation metrics

The results are summarized in Table 5, where best performances are highlighted in **bold**. Averages of performance metrics across all series are provided. MBB and MEB in the table stand for Moving

Blocks Bootstrap and Maximum Entropy Bootstrap, respectively.

Overall, the most accurate forecasts follow from combining the MEB algorithm for resampling and the Modified Ridge regularization routine as aggregation method. Using MEB for resampling and Modified LASSO regularization for aggregation is also competitive, as forecasts are more accurate than those from traditional benchmarks and the other *Bagging* approaches.

As anticipated, traditional regularization approaches are inferior, thus implying that the data generation processes of the forecasts should not be modified during validation and test. This is important, since regularization frameworks for forecasting have followed the traditional approach. Concerning the choice of forecasting method, regularized ensembles seem to benefit from ARIMA formulations. However, considering ensembles aggregated using the median, results from MEB.ARIMA are less competitive than those from MEB.ETS. Hence, forecasting replicas with ARIMA models may initially bring more variance to the ensemble, but this variance seems to be handled well by regularization.

Boxplots that summarize the performance of each method on each natural gas consumption time series are depicted in Figures 4 to 7. Overall, they are consistent with Table 5, with our Modified Regularization approaches (particularly Modified Ridge) generally outperforming comparators. Not only did Modified Regularization approaches presented considerably lower medians, but they were also less sensitive to outliers.

Figure 8 presents, for all countries, the differences between the natural gas consumption forecasts obtained through the MEB BaggedARIMA M-Ridge approach and the observed values throughout the test set period. The largest absolute differences between the forecasts and the real values occurs in Germany (DE). This can be largely attributed to variance in consumption, ranging from less than 200,000 TJ during the summer months to almost 500,000 TJ during the winter. Given this large variation, the RMSE for Germany is usually considered as outlier in the boxplots of Figure 4. In terms of MAPE, sMAPE and MASE (Figures 5, 6 and 7), the only country in which Modified Regularization approaches underperformed was Spain (ES). This country is an outlier for most methods. As illustrated in Figure 8, total consumption in Spain was considerably higher than expected, mainly due to a combination of two factors: the increasing use of natural gas, as opposed to coal, for electricity generation; and the growing industrial demand, which accounts for more than half of the country's total natural gas consumption (ENAGAS, 2020).

Resampling Algorithm	Forecast Approach	Combining Method	Average RMSE (TJ)	Average MAPE (%)	Average sMAPE (%)	Average MASE
Modified regularization approaches						
MEB	ETS	M-Ridge	8179.48	**9.42**	9.35	0.66
MEB	ETS	M-LASSO	8377.59	10.44	10.24	0.71
MEB	ARIMA	M-Ridge	**7568.67**	9.48	**9.32**	**0.62**
MEB	ARIMA	M-LASSO	8008.75	10.68	10.39	0.69
Traditional regularization approaches						
MEB	ETS	Ridge	10932.66	13.44	13.07	0.95
MEB	ETS	LASSO	11598.72	16.38	15.58	1.10
MEB	ARIMA	Ridge	10485.24	11.77	11.42	0.83
MEB	ARIMA	LASSO	10821.99	12.26	11.88	0.87
Traditional median aggregation						
MEB	ETS	Median	8597.17	12.92	11.61	0.74
MEB	ARIMA	Median	9318.01	14.12	12.86	0.79
Alternative Bagging approaches						
MBB	ETS	Median[a]	8702.57	11.31	10.87	0.73
MBB	ETS	BaggedCluster[b]	8608.12	11.24	10.75	0.73
MBB	ETS	BMC[c]	8717.61	11.41	11.01	0.73
Univariate time series forecasting benchmarks						
None	ETS	Single	8947.38	13.34	12.01	0.76
None	ARIMA	Single	9393.26	14.19	12.91	0.79
None	Additive HW	Single	9822.73	16.69	14.82	0.86
None	Multip. HW	Single	9363.76	11.03	10.97	0.75
None	Theta	Single	8876.91	11.64	11.10	0.74
Machine learning methods						
None	ANN	Single	12629.07	13.56	13.42	0.97
None	BC-ANN	Single	12894.53	14.02	13.11	0.93
None	SVR	Single	23755.28	20.16	22.10	1.68
None	RSSA	Single	55562.27	22.98	26.12	12.73
None	VSSA	Single	25160.11	22.15	24.11	4.73

Table 5: Forecast evaluation: Natural gas consumption over the months from January 2019 to December 2019, considering 12 steps ahead forecasts for all countries. Best methods in **bold**). Notes: [a, b, c] stand for the methods proposed in Bergmeir et al. (2016), Dantas & Cyrino Oliveira (2018) and Petropoulos et al. (2018), respectively. HW is the Holt-Winters Method.

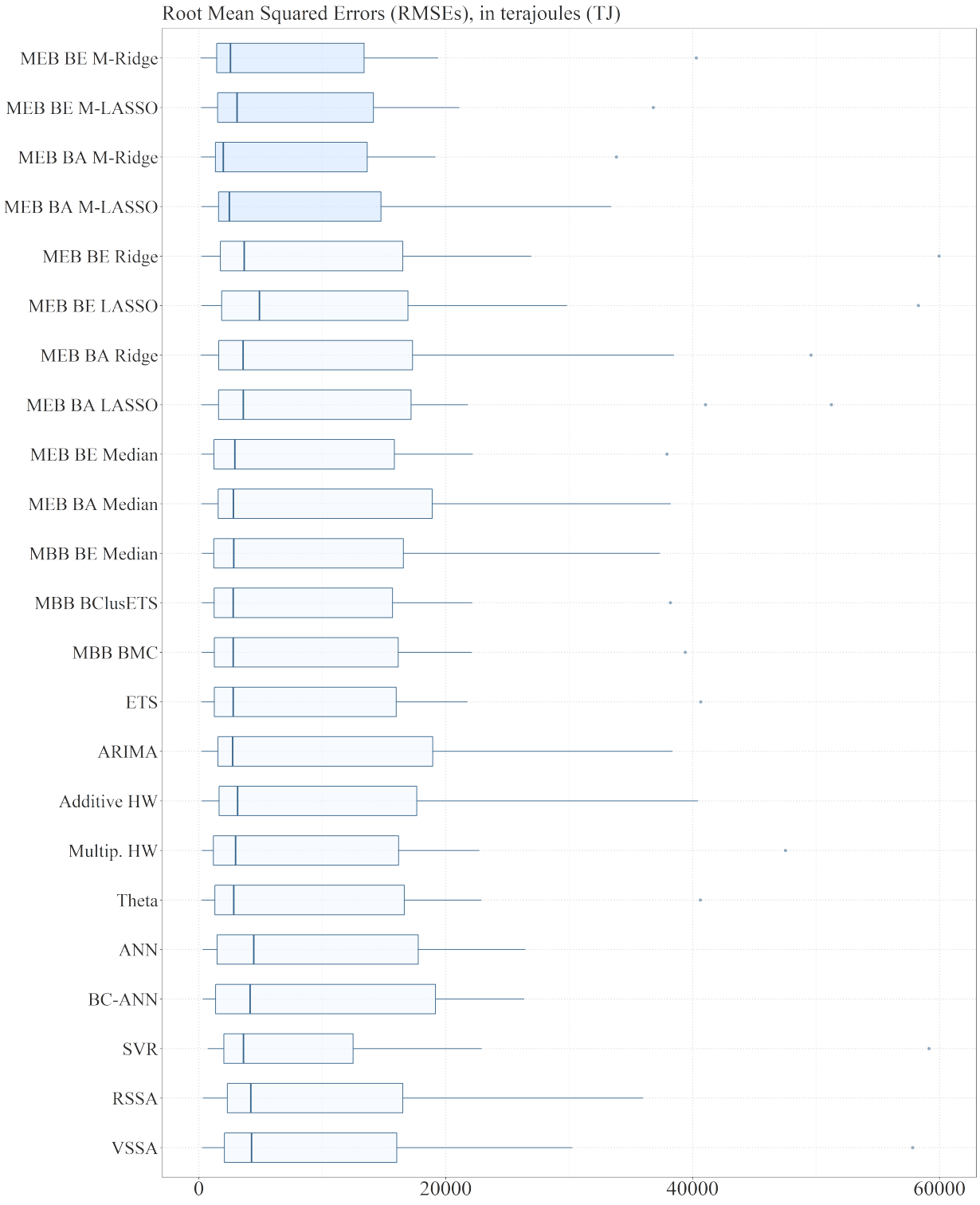

Figure 4: Boxplots – RMSE values (TJ) for each forecasting method considered. BE and BA stand for BaggedETS and BaggedARIMA.

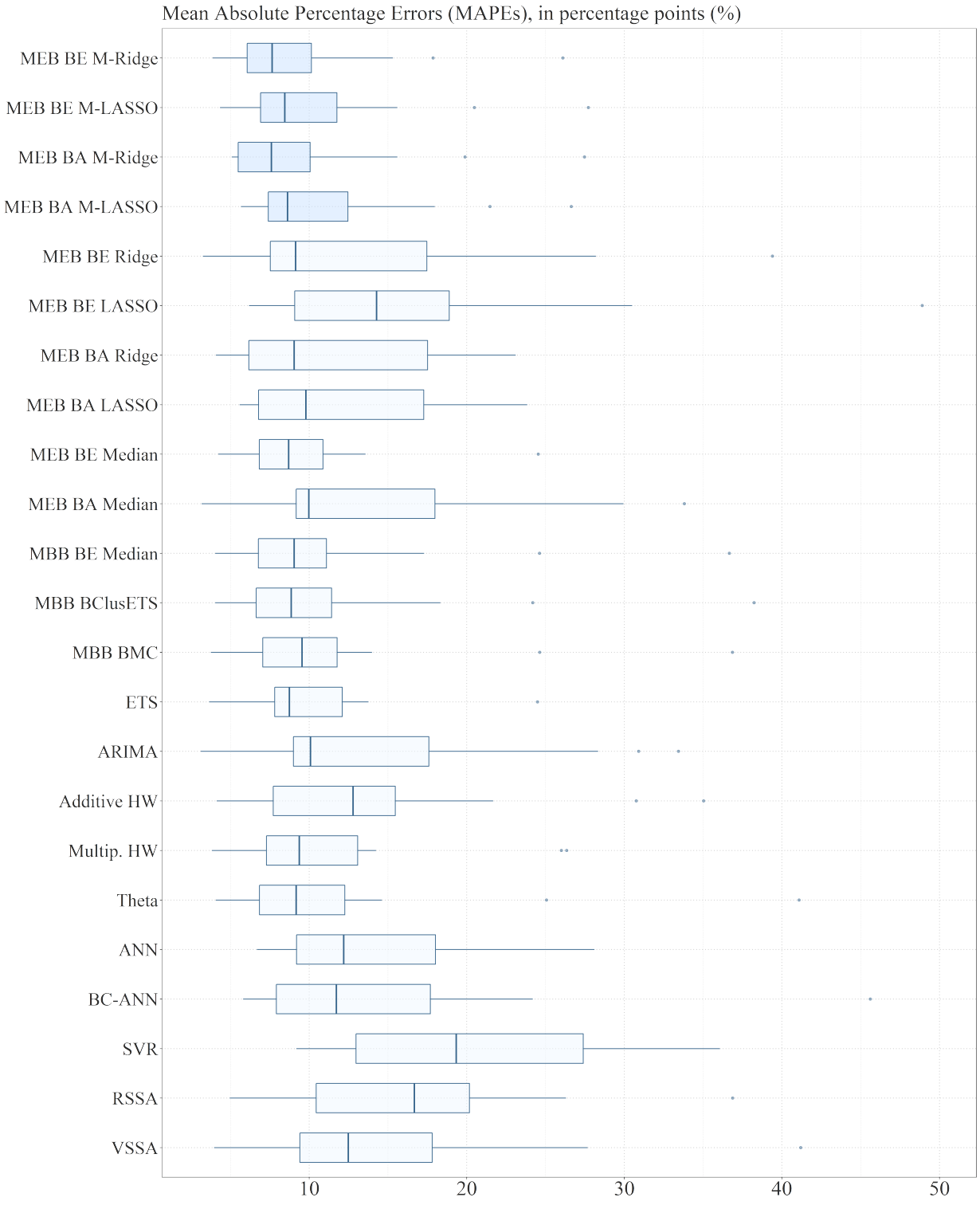

Figure 5: Boxplots – MAPE values (%) for each forecasting method considered.

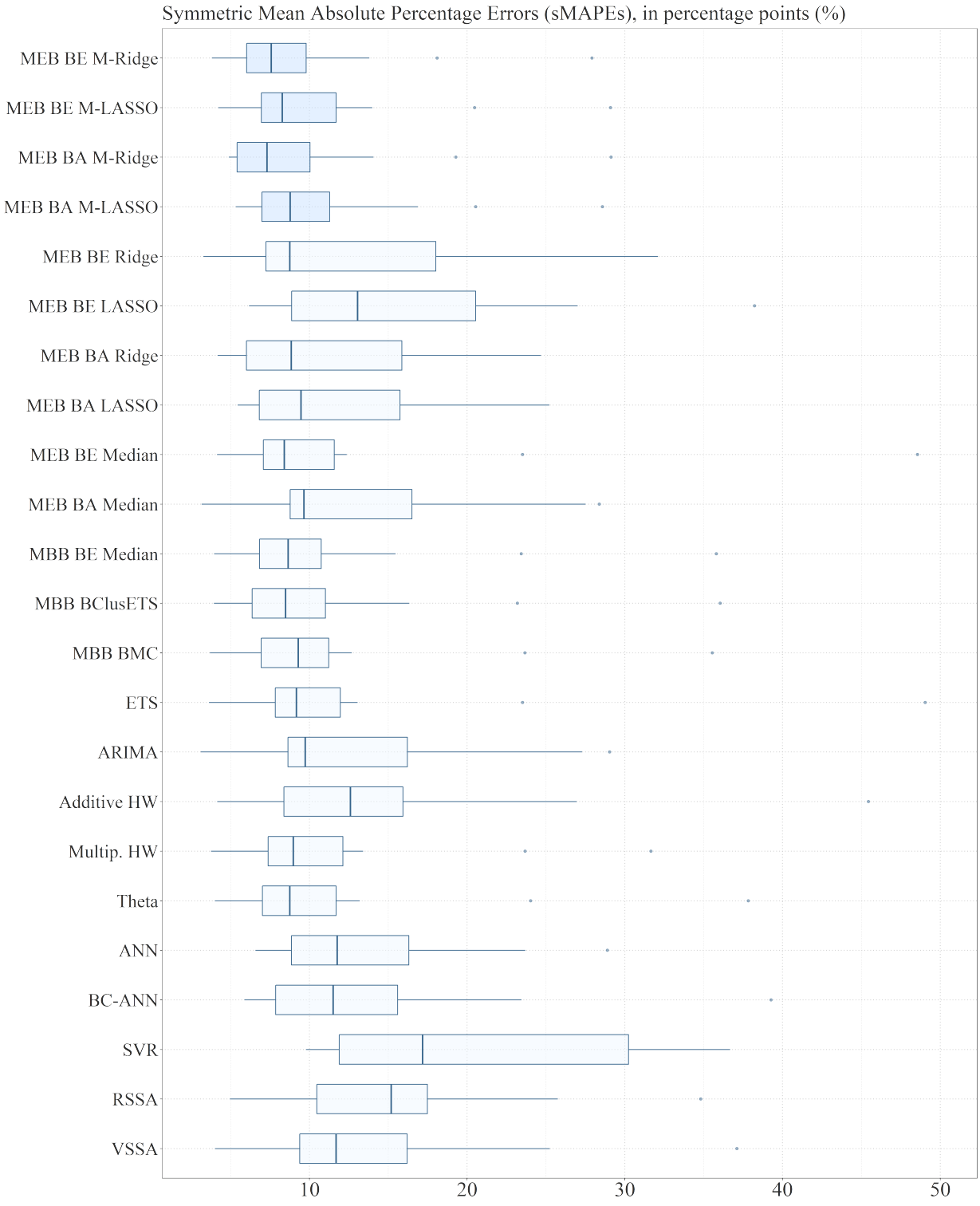

Figure 6: Boxplots – sMAPE values (%) for each forecasting method considered.

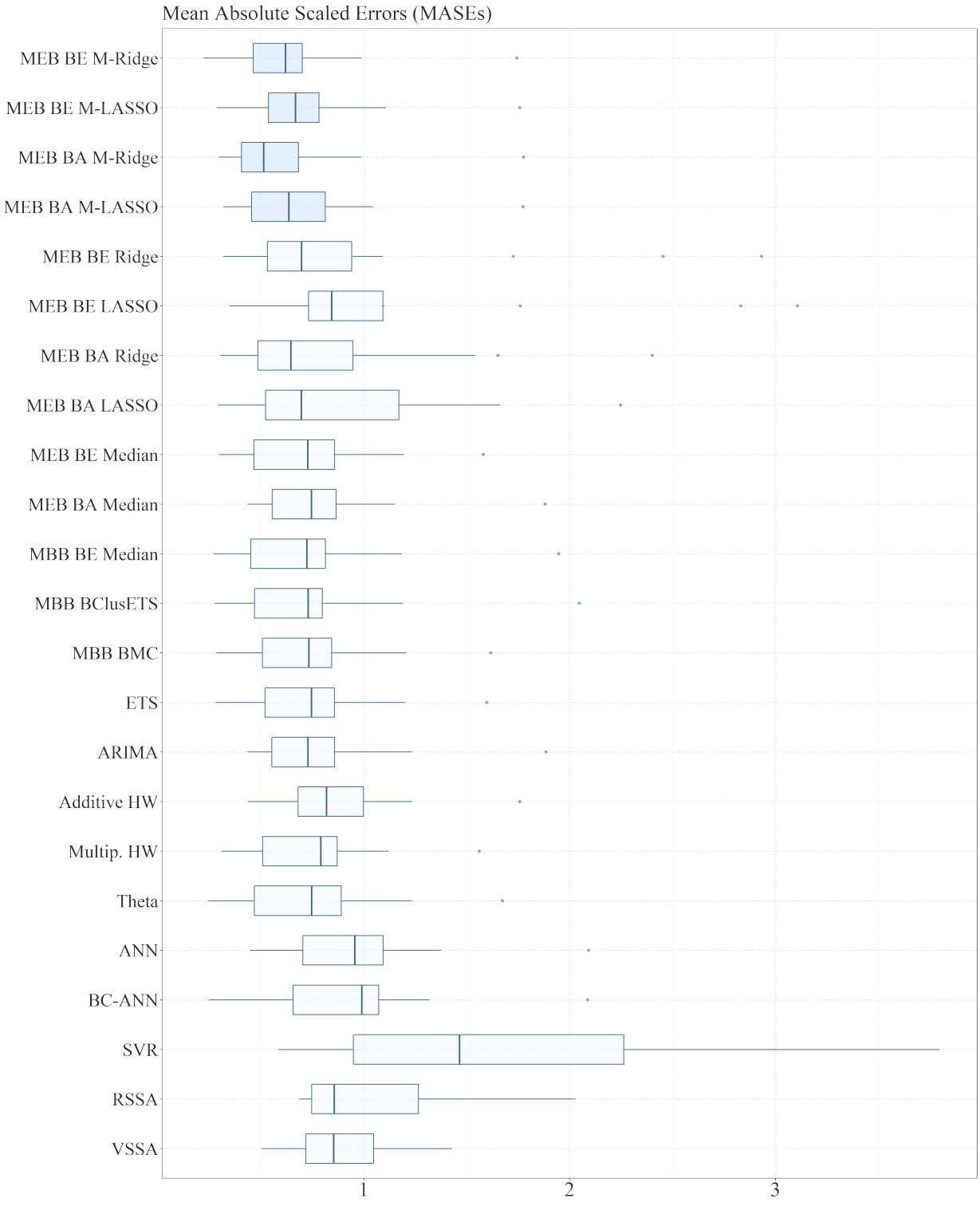

Figure 7: Boxplots – MASE values for each forecasting method considered.

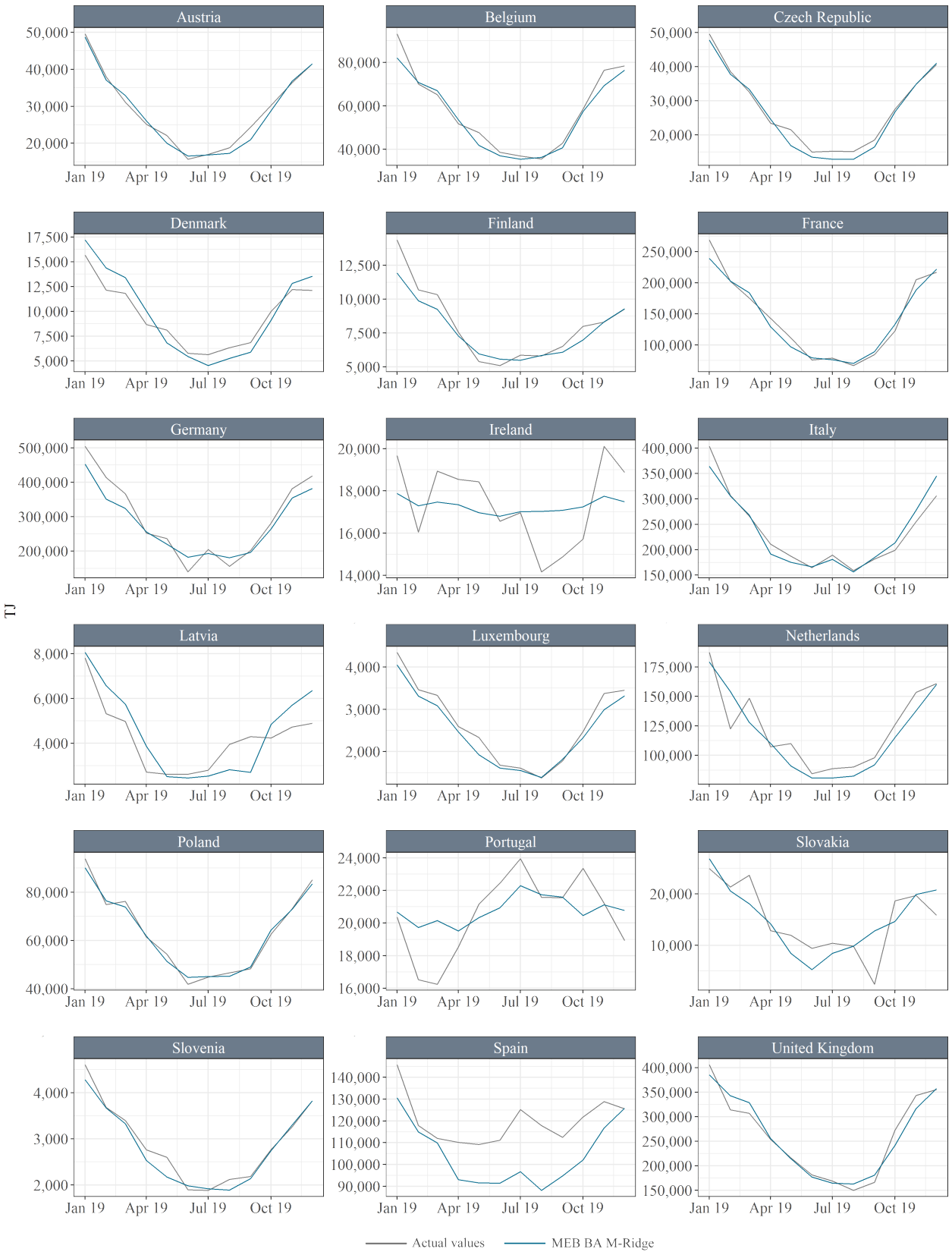

Figure 8: Out-of-sample forecasting: actual values in gray, MEB BaggedARIMA M-Ridge forecasts in blue.

5.2. Multiple comparisons with the best

MCB is a post-hoc, multiple comparison procedure that compares the average (across time series) rank of each method: statistically different performances are observed if the intervals of two methods do not overlap. The test is conducted using two set of methods: first, the average ranks of the four proposed approaches (MEB BaggedETS M-Ridge and M-LASSO and MEB BaggedARIMA M-Ridge and M-LASSO) are compared with those of competing *Bagging* approaches; then, the proposed approaches are compared with the statistical (time series forecasting) and machine learning benchmarks. This is done because the large number of approaches herein compared could lead to spurious interpretations.

The results from the MCB tests are depicted in Figures 9 and 10. They are in line with Table 5 and Figures 4 to 7, as the proposed approaches are shown to be very competitive. MEB BaggedARIMA M-Ridge and MEB BaggedETS M-Ridge stand out as the best methods in every comparison, with average ranks considerably lower than the comparators. Furthermore, the Modified Regularization ensembles (M-Ridge and M-LASSO) are the only methods that are statistically different than the worst method in every case. This can be observed, for instance, in the first chart of Figure 9, where all other competing *Bagging* algorithms, except for M-Ridge and M-LASSO, have average rank RMSEs similar to the average rank RMSE of the MEB BaggedETS LASSO (traditional regularization ensemble), which is the least competitive *Bagging* approach.

Among the statistical and machine learning benchmarks, the univariate Support Vector Regression (SVR) is the least competitive, followed by the Recurrent Singular Spectrum Analysis (RSSA). The Theta method, in turn, is the most accurate, although its average ranks are considerably higher than those from MEB BaggedARIMA M-Ridge and MEB BaggedETS M-Ridge.

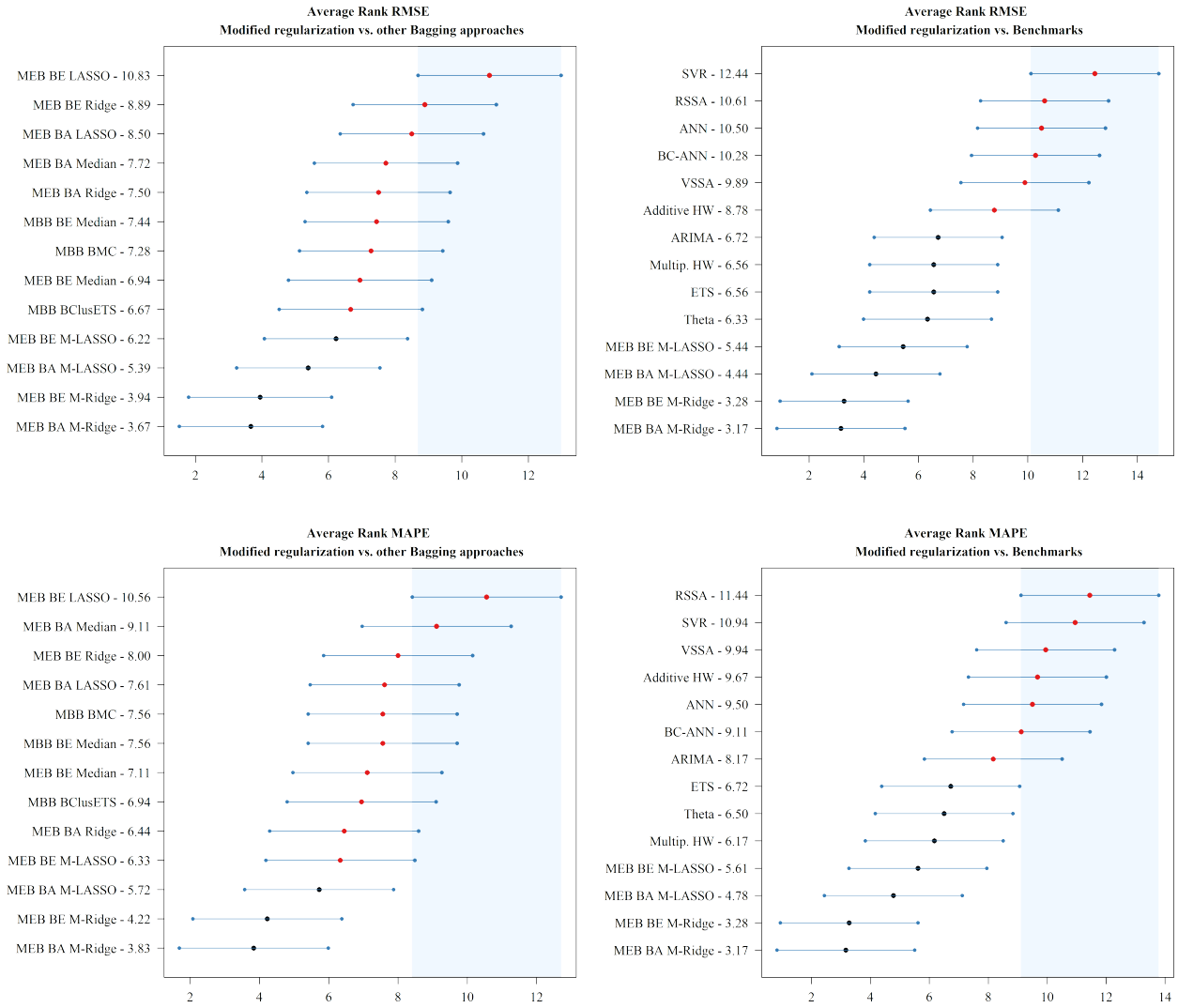

Figure 9: Multiple comparisons with the best for average rank RMSEs and MAPEs. Confidence bands at the 95% confidence level.

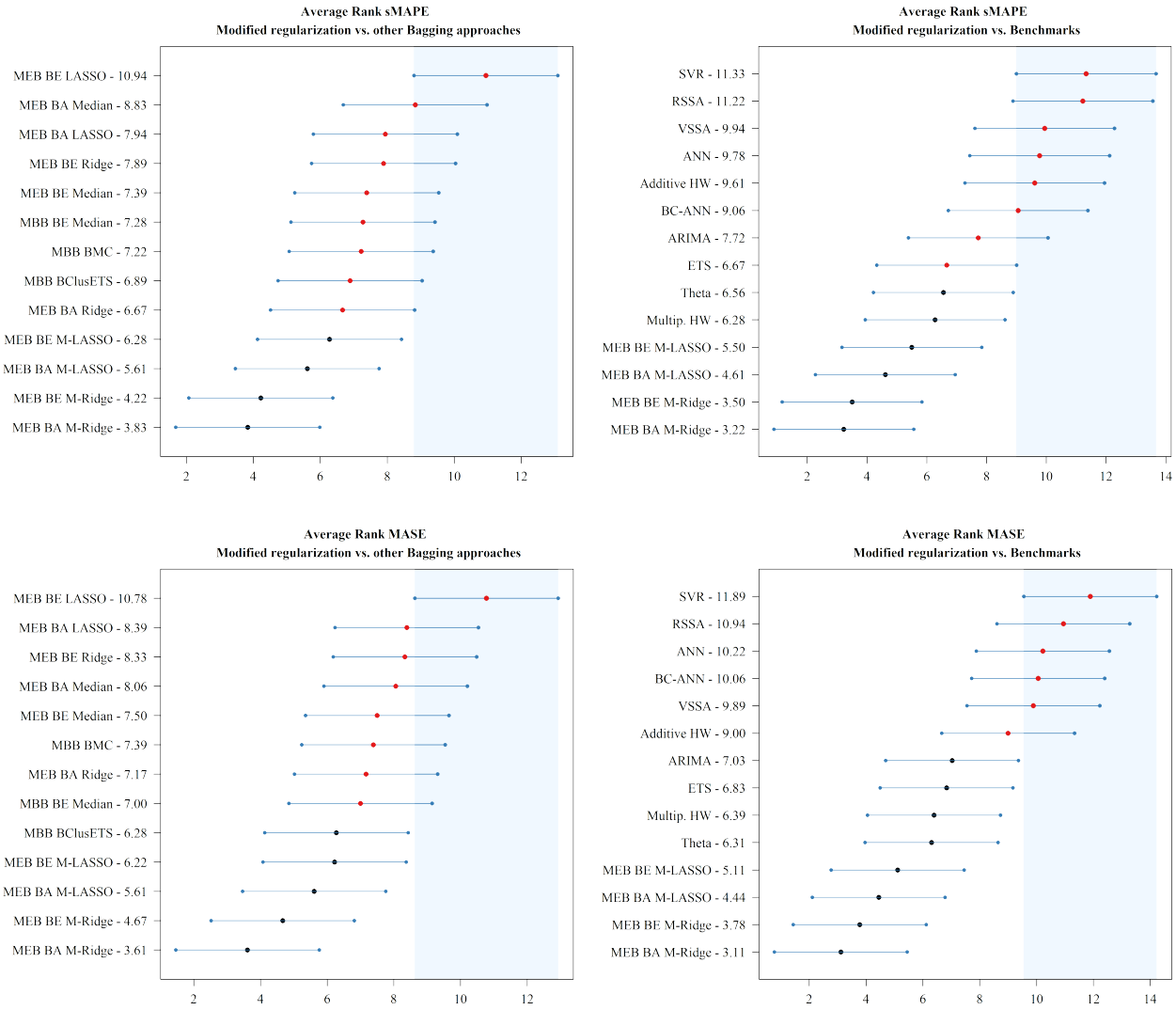

Figure 10: Multiple comparisons with the best for average rank sMAPEs and MASEs. Confidence bands at the 95% confidence level.

5.3. Sensitivity analysis and robustness checks

In this section, forecasting performance under alternative settings are considered. We begin by examining performance in a different forecasting period: June 2018 – May 2019. The results are summarized in Table 6. The relative performance across methods does not change much, with MEB.ARIMA.Modif.Ridge still providing the most accurate forecasts in terms of RMSE and MASE. Based on MAPE and sMAPE, however, the most accurate forecasts for this period result from using regularized ETS forecasts, with modified LASSO performing slightly better than modified Ridge.

Resampling Algorithm	Forecast Approach	Combining Method	Average RMSE (TJ)	Average MAPE (%)	Average sMAPE (%)	Average MASE
Modified regularization approaches						
MEB	ETS	M-Ridge	7598.44	8.27	8.23	0.57
MEB	ETS	M-LASSO	7627.39	**8.12**	**8.03**	0.56
MEB	ARIMA	M-Ridge	**6702.16**	9.30	8.91	**0.55**
MEB	ARIMA	M-LASSO	7148.78	9.62	9.12	0.57
Traditional regularization approaches						
MEB	ETS	Ridge	8875.20	12.12	11.09	0.74
MEB	ETS	LASSO	9170.33	13.60	11.97	0.79
MEB	ARIMA	Ridge	10000.05	13.33	12.52	0.84
MEB	ARIMA	LASSO	9905.76	11.86	12.01	0.84
Median aggregation						
MEB	ETS	Median	7845.66	11.53	10.21	0.63
MEB	ARIMA	Median	7273.48	13.81	10.99	0.63
Alternative Bagging approaches						
MBB	ETS	Median[a]	8090.62	9.85	9.30	0.61
MBB	ETS	BaggedCluster[b]	8132.64	9.56	9.11	0.61
MBB	ETS	BMC[c]	7941.03	10.43	9.69	0.62
Univariate time series forecasting benchmarks						
None	ETS	Single	8098.54	11.83	10.48	0.65
None	ARIMA	Single	7293.58	13.93	10.97	0.62
None	Additive HW	Single	8053.46	12.49	11.10	0.65
None	Multip. HW	Single	7946.58	9.94	9.51	0.62
None	Theta	Single	8146.22	12.29	10.96	0.68
Machine learning methods						
None	ANN	Single	10278.90	12.16	11.84	0.82
None	BC-ANN	Single	10363.49	12.44	11.88	0.84
None	SVR	Single	31811.02	47.62	33.90	2.30
None	RSSA	Single	10251.44	22.43	14.81	0.90
None	VSSA	Single	160202.18	27.23	24.04	6.69

Table 6: Forecast evaluation for the period between June 2018 and May 2019. Overall results (average of the evaluation metrics across all countries) considering 12 steps ahead forecasts (best in **bold**). Notes: See Table 5.

Different forecasting horizons are also considered: short run (steps 1–4), mid run (steps 5–8) and long run (steps 9–12). Forecasting performance is depicted in terms of average MASEs over each

selected horizon in Table 7. Overall, performance remains consistent, with modified regularized ensembles providing more accurate results across horizons.

Resampling Algorithm	Forecast Approach	Combining Method	Average MASE (steps 1–4)	Average MASE (steps 5–8)	Average MASE (steps 9–12)
Modified regularization approaches					
MEB	ETS	M-Ridge	0.73	0.60	0.65
MEB	ETS	M-LASSO	0.79	0.68	0.65
MEB	ARIMA	M-Ridge	**0.68**	**0.57**	**0.61**
MEB	ARIMA	M-LASSO	0.73	0.69	0.64
Traditional regularization approaches					
MEB	ETS	Ridge	1.04	0.84	0.96
MEB	ETS	LASSO	1.17	1.04	1.08
MEB	ARIMA	Ridge	0.93	0.80	0.77
MEB	ARIMA	LASSO	0.98	0.82	0.81
Median aggregation					
MEB	ETS	Median	0.79	0.78	0.65
MEB	ARIMA	Median	0.75	0.87	0.74
Alternative Bagging approaches					
MBB	ETS	Median[a]	0.79	0.76	0.65
MBB	ETS	BaggedCluster[b]	0.79	0.75	0.65
MBB	ETS	BMC[c]	0.78	0.76	0.64
Univariate time series forecasting benchmarks					
None	ETS	Single	0.83	0.80	0.65
None	ARIMA	Single	0.76	0.88	0.74
None	Additive HW	Single	0.80	1.01	0.76
None	Multip. HW	Single	0.82	0.71	0.71
None	Theta	Single	0.79	0.77	0.66
Machine learning methods					
None	ANN	Single	1.32	0.81	0.77
None	BC-ANN	Single	1.19	0.80	0.81
None	SVR	Single	1.58	2.10	1.35
None	RSSA	Single	2.47	7.10	28.61
None	VSSA	Single	1.42	2.96	9.81

Table 7: Average MASEs (best in **bold**) computed at different forecasting horizons (January 2019 – April 2019, May 2019 – August 2019, September 2019 – December 2019). Notes: See Table 5.

Potential differences between MEB and MBB in ensemble generation are also assessed, and results

are summarized in Table 8. It can be observed that modified regularization ensembles with MEB for resampling provide more accurate forecasts than those based on MBB. A possible explanation stems from how ensembles are created according to these two algorithms: MEB-generated ensembles are more diversified as MEB admits values near the original time series observations. This is an improvement over previous *Bagging* methods, especially as the MBB approach has been a standard benchmark for resampling monthly data (Petropoulos et al., 2018; Dantas & Cyrino Oliveira, 2018).

Resampling Algorithm	Forecast Approach	Combining Method	Average RMSE (TJ)	Average MAPE (%)	Average sMAPE (%)	Average MASE
Modified regularization approaches						
MEB	ETS	M-Ridge	8179.48	**9.42**	9.35	0.66
MEB	ETS	M-LASSO	8377.59	10.44	10.24	0.71
MEB	ARIMA	M-Ridge	**7568.67**	9.48	**9.32**	**0.62**
MEB	ARIMA	M-LASSO	8008.75	10.68	10.39	0.69
Traditional regularization approaches						
MEB	ETS	Ridge	10932.66	13.44	13.07	0.95
MEB	ETS	LASSO	11598.72	16.38	15.58	1.10
MEB	ARIMA	Ridge	10485.24	11.77	11.42	0.83
MEB	ARIMA	LASSO	10821.99	12.26	11.88	0.87
Modified regularization approaches using MBB for resampling						
MBB	ETS	M-Ridge	8304.18	9.45	9.34	0.66
MBB	ETS	M-LASSO	9174.36	11.26	11.00	0.76
MBB	ARIMA	M-Ridge	8242.21	9.92	9.74	0.66
MBB	ARIMA	M-LASSO	9648.41	11.64	11.31	0.77
Traditional regularization approaches using MBB for resampling						
MBB	ETS	Ridge	9677.88	11.26	11.01	0.77
MBB	ETS	LASSO	10720.74	14.37	13.80	0.94
MBB	ARIMA	Ridge	9472.95	11.23	11.42	0.87
MBB	ARIMA	LASSO	10719.51	13.05	12.52	0.89

Table 8: Comparisons with an alternative resampling algorithm (MBB). Overall results (average of the evaluation metrics across all countries) considering 12 steps ahead forecasts (best in **bold**). Block size for MBB comprises 24 observations.

5.4. Discussion and implications

The results outlined in Sections 5.1 to 5.3 endorse the strength of our proposed approaches. The study demonstrates the value of performing cross-validation and combination on the same set of forecasts (modified regularization approach), as the data generation process is kept, and predictions

are considerably better than when traditional regularization strategies are used. The performance gains are noteworthy since accurate forecasts are critical for profit/cost optimization and investment strategies, as well as for energy policies, whether in a regional or national scale. It should be noted that, for several countries, considerable variation in natural gas demand may be due to external factors, which cannot be captured by univariate forecasting methods, as for example gas on gas competition, markets expansions in access to key infrastructure, uncertainties over medium-term and long-term carbon pricing and emissions taxes inhibiting investment in gas infrastructure. Predictions could benefit from judgmental forecasts, possibly combining with quantitative methods, and thus leaving a question for future research: how to include experts' judgements into the ensemble?

An extension to a multivariate setting may be an alternative, particularly in short forecast horizons. For natural gas consumption forecasts, temperature is important when the forecast horizon is of a few days ahead. Carbon prices, in turn, are likely to become more important in the medium term, if used as a means to address climate change. The same holds for storage availability, as storage levels may influence prices, which in turn affect demand in mid term horizons. However, these are local variables for which data are more difficult to gain access. We hasten to add, however, that multivariate formulations usually fail to perform well when forecasting several steps ahead. This is because in most multivariate settings, the independent variables need to be previously or simultaneously forecasted so that their estimates are used to forecast the dependent variable. In a study like ours, variable selection would need to be conducted separately for each country involved, given variations in energy mix, local policies, and interactions between the different energy markets (e.g. various electricity and gas markets). A multivariate approach would add complexity, and may not lead to reliable forecasts or significant gains given the average errors obtained by the proposed ensembles for most time series. In this context, the combination of ensemble methods and univariate forecasting techniques is a promising alternative. In addition, such combination can be applied to a wide range of time series in different industries/sectors.

Another possibility for short-term predictions is the inclusion of different families of forecasting models in the forecast generation stage. For instance, hourly time series may exhibit three types of seasonality: a daily pattern, a weekly pattern, and an annual pattern. In such cases, the Trigonometric Exponential Smoothing State Space Model with Box-Cox Transformation, ARMA Errors, Trend and Seasonal Components (TBATS) approach by De Livera et al. (2011) is a promising alternative to be considered in the forecasting stage of the ensemble, given its consistent results when forecasting time series with multiple seasonal patterns. However, the TBATS approach is very computing intensive. Therefore, depending on the size of the ensemble (number of replicas), the

time availability and the computational power of the practitioner, using TBATS may be unfeasible. If this is the case, dynamic harmonic regressions with multiple seasonal periods (Hyndman & Athanasopoulos, 2021) may become attractive and can also be easily integrated in *Bagging*.

6. Summary and conclusions

This study proposed a novel, ensemble-based forecasting approach combining *Bagging* algorithms, time series methods and modified regularization techniques. In doing so, it integrates research from combining forecasts, statistics and committee learning machines. A Maximum Entropy Bootstrap (MEB) routine is adopted and a modified regularization approach allows for variable weighting schemes in the final stage of the ensemble.

As observed throughout the paper, the selected natural gas consumption time series differed considerably, thus highlighting the challenge of proposing a generalized method that provides reliable forecasts in every case considered. On these grounds, our proposal was confirmed to be a promising forecasting methodology: results and robustness checks demonstrated that the proposed ensemble offers accurate forecasts, whilst addressing different complex structures that are inherent to real world time series. Furthermore, the methodology is flexible and can be used to forecast time series of multiple frequencies and varying forecast horizons. Finally, we note that the use of the MEB procedure was shown to outperform the frequently used Moving Block Bootstrap (MBB) approach, which is common in forecasting ensembles. This is a contribution to the forecasting literature, as MBB has been the main benchmark for resampling monthly data under *Bagging*.

Further studies of energy consumption may benefit from a hierarchical disaggregation approach. For the natural gas sector, this would imply using the Decomposition and *Bagging* methods for each subsystem of the total consumption (Industrial, Electric Power, Residential, Transportation and Commercial). Such sector-tailored analysis may provide a more in-depth understanding of the demand for natural gas across different countries, potentially contributing to improve the forecasts.

Acknowledgements

This work was supported by the Brazilian Coordination for the Improvement of Higher Level Personnel (CAPES) under Grant [number 001]; the Brazilian National Council for Scientific and Technological Development (CNPq) under Grants [numbers 307403/2019-0 and 151079/2021-8]; and the Carlos Chagas Filho Research Support Foundation of the State of Rio de Janeiro (FAPERJ) under Grants [numbers 202.673/2018 and 211.086/2019].

References

Adeodato, P. J., Arnaud, A. L., Vasconcelos, G. C., Cunha, R. C., & Monteiro, D. S. (2011). MLP ensembles improve long term prediction accuracy over single networks. *International Journal of Forecasting*, *27*, 661–671. doi:10.1016/j.ijforecast.2009.05.029.

Agrawal, R. K., Muchahary, F., & Tripathi, M. M. (2019). Ensemble of relevance vector machines and boosted trees for electricity price forecasting. *Applied Energy*, *250*, 540–548. doi:10.1016/j.apenergy.2019.05.062.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*, 521–530. doi:10.1016/s0169-2070(00)00066-2.

Awajan, A. M., Ismail, M. T., & Wadi, S. A. (2018). Improving forecasting accuracy for stock market data using EMD-HW bagging. *PLOS ONE*, *13*, e0199582. doi:10.1371/journal.pone.0199582.

Azadeh, A., Asadzadeh, S., Saberi, M., Nadimi, V., Tajvidi, A., & Sheikalishahi, M. (2011). A neuro-fuzzy-stochastic frontier analysis approach for long-term natural gas consumption forecasting and behavior analysis: The cases of bahrain, saudi arabia, syria, and UAE. *Applied Energy*, *88*, 3850–3859. doi:10.1016/j.apenergy.2011.04.027.

Babatunde, O., Munda, J., & Hamam, Y. (2020). Power system flexibility: A review. *Energy Reports*, *6*, 101–106. doi:10.1016/j.egyr.2019.11.048.

Bai, Y., & Li, C. (2016). Daily natural gas consumption forecasting based on a structure-calibrated support vector regression approach. *Energy and Buildings*, *127*, 571–579. doi:10.1016/j.enbuild.2016.06.020.

Barak, S., & Sadegh, S. S. (2016). Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm. *International Journal of Electrical Power & Energy Systems*, *82*, 92–104. doi:10.1016/j.ijepes.2016.03.012.

Barbosa, S. M., Scotto, M. G., & Alonso, A. M. (2011). Summarising changes in air temperature over central europe by quantile regression and clustering. *Natural Hazards and Earth System Sciences*, *11*, 3227–3233. doi:10.5194/nhess-11-3227-2011.

Barrow, D. K., & Crone, S. F. (2016). Cross-validation aggregation for combining autoregressive neural network forecasts. *International Journal of Forecasting*, *32*, 1120–1137. doi:10.1016/j.ijforecast.2015.12.011.

Bastianin, A., Galeotti, M., & Polo, M. (2019). Convergence of european natural gas prices. *Energy Economics*, *81*, 793–811. doi:10.1016/j.eneco.2019.05.017.

BBC (2019). Gas heating ban for new homes from 2025. https://www.bbc.com/news/science-environment-47559920. Accessed: 2020-06-15.

Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and box–cox transformation. *International Journal of Forecasting*, *32*, 303–312. doi:10.1016/j.ijforecast.2015.07.002.

Beyca, O. F., Ervural, B. C., Tatoglu, E., Ozuyar, P. G., & Zaim, S. (2019). Using machine learning tools for forecasting natural gas consumption in the province of istanbul. *Energy Economics*, *80*, 937–949. doi:10.1016/j.eneco.2019.03.006.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211–252.

Box, G. E. P., & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco, CA, USA: Holden-Day, Inc.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. doi:10.1007/bf00058655.

Bühlmann, P. (1998). Sieve bootstrap for smoothing in nonstationary time series. *The Annals of Statistics*, *26*, 48–83. doi:10.1214/aos/1030563978.

CEU (2008). Council of the european union second strategic energy review: An eu energy security and solidarity action plan. http://aei.pitt.edu/39567/. Accessed: 2020-01-30.

Chen, J., Yu, J., Ai, B., Song, M., & Hou, W. (2019). Determinants of global natural gas consumption and import–export flows. *Energy Economics*, *83*, 588–602. doi:10.1016/j.eneco.2018.06.025.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, *6*, 3–73.

Cordeiro, C., & Neves, M. M. (2009). Forecasting time series with BOOT.EXPOS procedure. *REVSTAT – Statistical Journal*, *7*, 135–149.

Dantas, T. M., & Cyrino Oliveira, F. L. (2018). Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing. *International Journal of Forecasting*, *34*, 748–761. doi:10.1016/j.ijforecast.2018.05.006.

Dantas, T. M., Oliveira, F. L. C., & Repolho, H. M. V. (2017). Air transportation demand forecast through bagging holt winters methods. *Journal of Air Transport Management*, *59*, 116–123. doi:10.1016/j.jairtraman.2016.12.006.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*, 1513–1527. doi:10.1198/jasa.2011.tm09771.

De Oliveira, E. M., & Cyrino Oliveira, F. L. (2018). Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. *Energy*, *144*, 776–788. doi:10.1016/j.energy.2017.12.049.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26. doi:10.1214/aos/1176344552.

ENAGAS (2020). Enagas annual report 2019. https://www.enagas.es/WEBCORP-static/Informe_Anual_2019/. Accessed: 2020-07-15.

EUROSTAT (2020). European Statistics supply of gas – gross inland consumption – monthly data. https://ec.europa.eu/eurostat/web/energy/data/database. Accessed: 2020-06-02.

Fan, G.-F., Guo, Y.-H., Zheng, J.-M., & Hong, W.-C. (2020). A generalized regression model based on hybrid empirical mode decomposition and support vector regression with back-propagation neural network for mid-short-term load forecasting. *Journal of Forecasting*, *39*, 737–756. doi:10.1002/for.2655.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, *121*, 256–285. doi:10.1006/inco.1995.1136.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*. doi:10.18637/jss.v033.i01.

Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. (2001). *Analysis of time series structure: SSA and related techniques*. (1st ed.). Boca Raton: Chapman & Hall/CRC.

Goodwin, P. (2010). The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong. *Foresight: The International Journal of Applied Forecasting*, (pp. 30–33).

Hall, M. A. (1999). Correlation-based feature selection for machine learning. https://www.cs.waikato.ac.nz/~mhall/. Accessed: 2021-06-21.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67. doi:10.1080/00401706.1970.10488634.

Holt, C. C. (1957). *Forecasting seasonals and trends by exponentially weighted moving averages*. ONR Memorandum, vol. 52. Pittsburgh: PA7 Carnegie Institute of Technology.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*, 5–10. doi:10.1016/j.ijforecast.2003.09.015.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *454*, 903–995. doi:10.1098/rspa.1998.0193.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2021). *forecast: Forecasting functions for time series and linear models*. URL: http://pkg.robjhyndman.com/forecast R package version 8.15.

Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer Berlin Heidelberg. URL: https://doi.org/10.1007/978-3-540-71918-2. doi:10.1007/978-3-540-71918-2.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice*. (3rd ed.). OTexts: Melbourne, Australia. URL: OTexts.com/fpp3.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*. doi:10.18637/jss.v027.i03.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688. doi:10.1016/j.ijforecast.2006.03.001.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454. doi:10.1016/s0169-2070(01)00110-8.

IEA (2020). World energy outlook 2020. https://www.iea.org/reports/world-energy-outlook-2020. Accessed: 2021-07-18.

IEA (2021). Global energy review 2021. https://www.iea.org/reports/global-energy-review-2021/natural-gas. Accessed: 2021-07-19.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge, & editor (Eds.), *Reports of the Faculty of Mathematics and Informatics. Delft University of Technology* (p. 405–416). North Holland / Elsevier.

Khwaja, A., Zhang, X., Anpalagan, A., & Venkatesh, B. (2017). Boosted neural networks for improved short-term electric load forecasting. *Electric Power Systems Research*, *143*, 431–437. doi:10.1016/j.epsr.2016.10.067.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. (2005). The m3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*, 397–409. doi:10.1016/j.ijforecast.2004.10.003.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*, 1217–1241. doi:10.1214/aos/1176347265.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*, 159–178. doi:10.1016/0304-4076(92)90104-Y.

Liebensteiner, M., & Wrienz, M. (2020). Do intermittent renewables threaten the electricity supply security? *Energy Economics*, *87*, 104499. doi:10.1016/j.eneco.2019.104499.

Ma, T., Wang, C., Wang, J., Cheng, J., & Chen, X. (2019). Particle-swarm optimization of ensemble neural networks with negative correlation learning for forecasting short-term wind speed of wind farms in western china. *Information Sciences*, *505*, 157–182. doi:10.1016/j.ins.2019.07.074.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153. doi:10.1002/for.3980010202.

Makridakis, S., & Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476. doi:10.1016/s0169-2070(00)00057-1.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*, 802–808. doi:10.1016/j.ijforecast.2018.06.001.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2019). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, . doi:10.1016/j.ijforecast.2019.04.014.

Matijaš, M., Suykens, J. A. K., & Krajcar, S. (2013). Load forecasting using a multivariate meta-learning system. *Expert Systems with Applications*, *40*, 4427–4437. doi:10.1016/j.eswa.2013.01.047.

Meira, E., Cyrino Oliveira, F. L., & De Menezes, L. M. (2021a). Point and interval forecasting of electricity supply via pruned ensembles. *Energy*, *232*, 121009. doi:10.1016/j.energy.2021.121009.

Meira, E., Oliveira, F. L. C., & Jeon, J. (2021b). Treating and pruning: New approaches to forecasting model selection and combination using prediction intervals. *International Journal of Forecasting*, *37*, 547–568. doi:10.1016/j.ijforecast.2020.07.005.

Nock, R., & Gascuel, O. (1995). On learning decision committees. In *Machine Learning Proceedings 1995* (pp. 413–420). Elsevier. doi:10.1016/b978-1-55860-377-6.50058-x.

OPEC (2020). World oil outlook 2020. https://woo.opec.org/index.php. Accessed: 2021-07-19.

Panapakidis, I. P., & Dagoumas, A. S. (2017). Day-ahead natural gas demand forecasting based on the combination of wavelet transform and ANFIS/genetic algorithm/neural network model. *Energy*, *118*, 231–245. doi:10.1016/j.energy.2016.12.033.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., & et al. (2020). Forecasting: theory and practice. arXiv:2012.03854.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*, 545–554. doi:10.1016/j.ejor.2018.01.045.

Petropoulos, F., Wang, X., & Disney, S. M. (2019). The inventory performance of forecasting methods: Evidence from the m3 competition data. *International Journal of Forecasting*, *35*, 251–265. doi:10.1016/j.ijforecast.2018.01.004.

Potočnik, P., Soldo, B., Šimunović, G., Šarić, T., Jeromen, A., & Govekar, E. (2014). Comparison of static and adaptive models for short-term residential natural gas forecasting in croatia. *Applied Energy*, *129*, 94–103. doi:10.1016/j.apenergy.2014.04.102.

Quinlan, J. R. (1996). Bagging, boosting, and c4.s. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1* AAAI'96 (pp. 725–730). AAAI Press. URL: http://dl.acm.org/citation.cfm?id=1892875.1892983.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Rendon-Sanchez, J. F., & de Menezes, L. M. (2019). Structural combination of seasonal exponential smoothing forecasts applied to load forecasting. *European Journal of Operational Research*, *275*, 916 – 924. doi:10.1016/j.ejor.2018.12.013.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. https://apps.dtic.mil/sti/pdfs/ADA164453.pdf. Accessed: 2021-06-22.

Sánchez-Úbeda, E. F., & Berzosa, A. (2007). Modeling and forecasting industrial end-use natural gas consumption. *Energy Economics*, *29*, 710–742. doi:10.1016/j.eneco.2007.01.015.

Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2019). Improving the forecasting performance of temporal hierarchies. *PLOS ONE*, *14*, e0223422. doi:10.1371/journal.pone.0223422.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430. doi:10.1002/for.928.

Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7*, 13–26. doi:10.1080/03610927808827599.

Sun, S., Sun, Y., Wang, S., & Wei, Y. (2018). Interval decomposition ensemble approach for crude oil price forecasting. *Energy Economics*, *76*, 274–287. doi:10.1016/j.eneco.2018.10.015.

Szafranek, K. (2019). Bagged neural networks for forecasting polish (low) inflation. *International Journal of Forecasting*, *35*, 1042–1059. doi:10.1016/j.ijforecast.2019.04.007.

Taşpınar, F., Çelebi, N., & Tutkun, N. (2013). Forecasting of daily natural gas consumption on regional basis in turkey using various computational methods. *Energy and Buildings*, *56*, 23–31. doi:10.1016/j.enbuild.2012.10.023.

Theodosiou, M. (2011). Forecasting monthly and quarterly time series using STL decomposition. *International Journal of Forecasting*, *27*, 1178–1195. doi:10.1016/j.ijforecast.2010.11.002.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. (1st ed.). New York: Springer.

Vinod, H. (2004). Ranking mutual funds using unconventional utility theory and stochastic dominance. *Journal of Empirical Finance*, *11*, 353–377. doi:10.1016/j.jempfin.2003.06.002.

Vinod, H. D. (2006). Maximum entropy ensembles for time series inference in economics. *Journal of Asian Economics*, *17*, 955–978. doi:10.1016/j.asieco.2006.09.001.

Vinod, H. D., & López-de-Lacalle, J. (2009). Maximum entropy bootstrap for time series: The meboot R Package. *Journal of Statistical Software*, *29*. doi:10.18637/jss.v029.i05.

Vondráček, J., Pelikán, E., Konár, O., Čermáková, J., Eben, K., Malý, M., & Brabec, M. (2008). A statistical model for the estimation of natural gas consumption. *Applied Energy*, *85*, 362–370. doi:10.1016/j.apenergy.2007.07.004.

Wang, B., & Wang, J. (2020). Energy futures and spots prices forecasting by hybrid SW-GRU with EMD and error evaluation. *Energy Economics*, *90*, 104827. doi:10.1016/j.eneco.2020.104827.

Webb, G. I. (2000). Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, *40*, 159–196. doi:10.1023/A:1007659514849.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*, 324–342. doi:10.1287/mnsc.6.3.324.

Wood, D. A. (2016). Natural gas imports to europe: The frontline of competition between LNG and pipeline supplies. *Journal of Natural Gas Science and Engineering*, *36*, A1–A4. doi:10.1016/j.jngse.2016.09.065.

Yalta, A. T. (2011). Analyzing energy consumption and GDP nexus using maximum entropy bootstrap: The case of Turkey. *Energy Economics*, *33*, 453–460. doi:10.1016/j.eneco.2010.12.005.

Zhang, J.-L., Zhang, Y.-J., & Zhang, L. (2015). A novel hybrid method for crude oil price forecasting. *Energy Economics*, *49*, 649–659. doi:10.1016/j.eneco.2015.02.018.

Özmen, A., Yılmaz, Y., & Weber, G.-W. (2018). Natural gas consumption forecast with MARS and CMARS models for residential users. *Energy Economics*, *70*, 357–381. doi:10.1016/j.eneco.2018.01.022.